UNIVERSITY OF CALIFORNIA

Santa Barbara

Developing experimental and computational single-cell sequencing techniques to study complex mammalian and bacterial systems

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Chemical Engineering

by

Chatarin Wangsanuwat

Committee in charge:

Professor Siddharth Dey, Chair

Professor Michelle O'Malley

Professor Arnab Mukherjee

Professor Ryan Stowers

September 2021

The dissertation of Chatarin Wangsanuwat is approved.

_____

Michelle O'Malley

_____

Arnab Mukherjee

_____

Ryan Stowers

_____

Siddharth Dey, Committee Chair

August 2021

Developing experimental and computational single-cell techniques to uncover biological insights

DEDICATION

This dissertation is dedicated to my grandfather, Prakob Katanyutanon

ACKNOWLEDGEMENTS

I would like to extend my gratitude towards Dr. Siddharth Dey for his guidance and support during my Ph.D. program. I learned and grew tremendously as a scientist through his continual challenge for me to think more critically and thoroughly. I would like to thank my committee members, Dr. Michelle O'Malley, Dr. Arnab Mukherjee, Dr. Ryan Stowers, and Dr. Irene Chen, for their support and insightful feedbacks.

I would also like to express deep gratitude to my parents, Pakinee and Chairode Wangsanuwat, for their unwavering faith in me and unconditional love for me, and to my favorite brother, Pakawat Wangsanuwat, for being an amazing big brother and taking care of me. I would like to also thank all my friends, old and new, who have been with me to enjoy all the good times and commiserate all the tough times throughout this Ph.D. journey of mine.

From the bottom of my heart, thank you.

VITA OF CHATARIN WANGSANUWAT
Aug 2021

EDUCATION

Bachelor of Science in Chemical and Biological, Princeton University, May 2016 (magna cum laude)
Doctor of Philosophy in Chemical Engineering, University of California, Santa Barbara, August 2021 (expected)

PROFESSIONAL EMPLOYMENT

2017-20: Teaching Assistant, Department of Chemical Engineering, University of California, Santa Barbara

PUBLICATIONS

Wangsanuwat C*, Heom KA*, Liu E, O'Malley MA, Dey SS. *Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion, BMC Genomics* 21, 717 (2020). https://doi.org/10.1186/s12864-020-07134-4. (*s*hared first authorship).*

Wangsanuwat C, Chialastri AJ, Aldeguer JF, Rivron NC, Dey SS. *A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing, Cell Reports Methods* (2021). https://doi.org/10.1016/j.crmeth.2021.100060.

AWARDS

CSP Technologies Teacher-Scholar Fellowship, Department of Chemical Engineering, University of California, Santa Barbara, 2019

FIELDS OF STUDY

Major Field: Chemical engineering

Emphasis in Bioengineering

ABSTRACT

Developing experimental and computational single-cell sequencing techniques to study complex mammalian and bacterial systems

by

Chatarin Wangsanuwat

Single-cell genomics is a rapidly advancing field that is leading to unprecedented new insights into complex biological systems, ranging from diverse microbial populations to early mammalian embryogenesis. This dissertation has contributed to the field by presenting techniques that can used to study cell differentiation and tissue development: 1. a mathematical model to reconstruct cellular lineage in pre-implantation embryos, 2. an efficient mRNA enrichment method for prokaryotic cells, which can also be used to enrich for mammalian non-coding RNA (ncRNA).

Cellular lineage reconstruction: Lineage reconstruction is central to understanding tissue development and maintenance. While various tools to infer cellular relationships have been established, these methods typically involve genetic modification and have a clonal resolution. This dissertation introduced scPECLR, a probabilistic algorithm to endogenously infer lineages at a single cell-division resolution using the epigenetic mark 5-hydroxymethylcytosine (5hmC) in single cells. When applied to 8-cell mouse embryos, scPECLR predicted the full lineage trees with greater than 95% accuracy. Furthermore, a protocol to detect both 5hmC and genomic DNA from the same single cell was developed. Information from genomic DNA, in combination with scPECLR, could allow us to identify cellular lineages more accurately and expand the reconstruction to even larger trees. The

high accuracy of scPECLR, using only endogenous marks, suggests that the method can be directly extended to study human development.

Low-input bacterial mRNA sequencing: RNA sequencing is a powerful approach to quantify the genome-wide distribution of mRNA molecules in a population to gain understanding of cellular functions and phenotypes. However, compared to mammalian cells, mRNA sequencing of bacterial samples is more challenging due to their 100-fold lower RNA quantities and the absence of a poly(A)-tail that typically enables enrichment of mRNA. To overcome these limitations, an effective mRNA enrichment method called EMBR-seq was introduced. The method resulted in greater than 90% of the sequenced *E. coli* RNA reads deriving from mRNA, which originally contributed to lower than 5% of total RNA in a cell. Moreover, EMBR-seq successfully quantified mRNA from 20 picogram total RNA, a level 500-fold lower than required in existing commercial kits. In addition, EMBR-seq can be combined with an orthogonal rRNA depletion method, RNase H, to improve the efficiency of mRNA enrichment. Due to its simplicity and efficiency, EMBR-seq could potentially be extended to a single-cell resolution to advance developments in bacterial mRNA sequencing and to investigate gene expression patterns in non-model microbial species.

Mammalian non-coding RNA sequencing: Despite being highly investigated, mRNA accounts for less than 5% of the mammalian RNA that are transcribed. There are many other types of non-coding RNAs capable of performing various functions, including transcriptional regulation. Similar to bacterial mRNA, mammalian ncRNA lacks a poly(A)-tail. EMBR-seq was shown to deplete mammalian rRNA and increase the number of unique ncRNA detected.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

After an egg is fertilized, the one single cell of fertilized egg will divide and differentiate into many cells capable of performing different functions. The process where an organism develops from a single-cell zygote to a multi-cellular organism is complex and well-regulated. However, how the first two distinct cell types emerge in an early embryo is not well-understood. In order to study cell differentiation and tissue development, there are two sets of information required: relationships between the cells and cell types. To gain this information, we need the tools to identify cellular lineages and cell types at single cells.

A. *Tissue development and cell fate determination*

In developmental biology field, one central question is how a particular cell develops into its final cell type, a process broadly known as cell fate determination. Each cell in a developing embryo receives molecular signals from neighboring cells in the form of proteins, RNAs, and surface interactions. Animals undergo a similar conserved series of events called embryogenesis during their early development (Saenko et al., 2008). There is a basic set of the same proteins and mRNAs involved in embryogenesis across species. This evolutionary conservation is one of the reasons why model organisms such as fruit fly (*Drosophila melanogaster*), zebrafish (*Danio rerio*), and mouse (*Mus musculus*) can be used to study human embryogenesis and development. Mouse model in particular is the most commonly used animal model to study human development and diseases because: 1. As mammals, mice are biologically very similar to humans; 2. Mice can be genetically manipulated to mimic human diseases and can be inbred to preserve their genetic information; 3. They have a shorter lifespan and thus are cost-effective.

Understanding the process of tissue development and cell-fate determination in humans has applications in regenerative biology and clinical medicine. The knowledge of how one cell type divides into another cell type would potentially allow us to genetically engineer one stem cell type to any cell type of interest in a systematic manner. Moreover, the knowledge of how early human embryos develop could help improve the success of *in vitro* fertilization (IVF), by allowing researchers to perform preimplantation genetic diagnosis with minimal embryo damage. In order to gain insights into the cell differentiation process, cellular lineage information needs to be obtained.

B. *Lineage tracing and lineage reconstruction*

Lineage tracing, lineage tracking, or lineage reconstruction is an experimental method to map the fates of cells in a system. Prospective lineage analysis introduces lineage tracers and follows cells forward in development. On the other hand, retrospective lineage analysis identifies a tracer or marker and uses that information to infer past developmental relationship without experimental intervention.

1. Lineage tracing methods using fluorescent protein

There are many lineage tracing methods currently available, most of which involves a reporter transgene. One major category is the use of a fluorescent protein. DNA plasmid containing reporter transgenes such as green fluorescent protein (GFP) can be transfected into cells. The cells expressing GFP can then be traced. Since the plasmid is not integrated into the genome of the progenitor, it becomes diluted after a series of cell division and fails to label the entire lineage (LoTurco et al., 2009). A similar method using retrovirus containing reporter transgenes improves upon the dilution limitation and has been used in

vertebrate animal models to identify clonal relationships (Frank and Sanes, 1991; Turner and Cepko, 1987). Another popular method is genetic recombination using Cre-*loxP*. In Cre-*loxP* system, mice are engineered to express Cre recombinase with a reporter transgene under the control of a promoter. In cells that express Cre recombinase, the reporter transgene is expressed. Moreover, multicolor tracing is possible under some mouse lines, including Brainbow and Confetti (Livet et al., 2007; Snippert et al., 2010). A more recent method called MEMOIR (memory by engineered mutagenesis with optical *in situ* readout) uses a set of barcoded recording elements called scratchpads that are altered by CRISPR-Cas9. These scratchpads are subsequently recorded by single-molecule RNA fluorescence hybridization (smFISH) (Frieda et al., 2017). Unlike previous methods that infer clones, MEMOIR allows lineage tracing at a single cell-division resolution in mouse embryonic stem cells for up to 3 cell divisions (Frieda et al., 2017).

While florescent protein has allowed lineage tracing and provided many biological insights, three major general limitations exist: 1. these methods require genetic modification, preventing the methods to be extended to directly study human cells; 2. They are microscopy-based, which would require the organisms to be translucent, such that the fluorescence can be tracked; 3. These methods generally yield only clonal resolution, which cannot be used to study cell differentiation, as lineages at a single cell-division resolution is required. Furthermore, with advances in sequencing technology in the last decades, more powerful lineage tracing methods utilizing next-generation sequencing have been developed.

2. Lineage tracing methods using next-generation sequencing

During the last few decades, there has been an unimaginable improvement in the ability to swiftly and accurately determine nucleic acid sequences. Moreover, novel approaches and

improvements of existing methods have significantly reduced the cost and increased the throughput. The Human Genome Project's first human genome sequence (~ 3 billion bases) took 13 years to complete (1990-2003) and costed $2.7 billion (Wetterstrand). In 2005, the first "next-generation" sequencing (NGS) technology became available. With NGS technology, sequencing 6 billion bases of human genome costed ~$14 million in 2006, and by late 2015, the cost had fallen below $1500 (Wetterstrand).

Along with the advances of sequencing technology, new tools to measure the sequences of individual cells, collectively called single-cell sequencing, have emerged. The process generally involves isolating a single cell, amplifying the whole genome, region of interest, and/or mRNA, constructing sequencing libraries, and applying next-generation sequencing. Compared with bulk sequencing, which can only measure the average of many cells, single-cell technologies can detect heterogeneity among individual cells (Wen and Tang, 2018), identify rare cell types, and show evolutionary relationships of various cells. The drop in sequencing cost and single-cell sequencing technologies have led to many biological insights and has enabled previously-unavailable scientific investigation, including in the field of lineage tracing.

CRIPR-Cas9 genome-editing technology has been used in combination with next-generation sequencing to track and reconstruct lineage relationships in complex, multicellular organisms. GESTALT (genome editing of synthetic target arrays for lineage tracing) is a pioneering method that utilizes CRISPR-Cas9 technology (McKenna et al., 2016). It uses barcodes that consist of multiple CRISPR-Cas9 target sites. These barcodes progressively accumulate unique mutations over cell divisions and can be recovered using sequencing (McKenna et al., 2016). Cellular relationships are determined based on the

shared mutation among the cells. Improving upon this idea, ScarTrace, LINNAEUS (lineage tracing by nuclease-activated editing of ubiquitous sequences), and another method use Cas9 technology to simultaneously identify clonal history and cell types of thousands of cells from different adult zebrafish's organs (Alemany et al., 2018; Raj et al., 2018; Spanjaard et al., 2018). Similarly, another method uses homing CRISPR guide RNA (hgRNA) to create a substantially larger mutation sites, which allows lineage tracing of a whole mouse (Kalhor et al., 2017). While CRIPS-Cas9 methods have greatly improved the reconstruction power to whole organism and eliminated the microscopy requirement, they still require a cell line capable of carrying the transgenes to be created and cannot yield single cell-division resolution.

3. Lineage reconstruction using endogenous mutation

Advances in single-cell sequencing also have enabled naturally occurring mutations as tools to infer cell lineages retrospectively. Similar to transgene introduction, somatic mutations mark the progeny of the dividing cells where the mutations occur (Salipante et al., 2010). The genomic variations used, from least to most frequently mutated, are retrotransposons, copy-number variants (CNVs), single-nucleotide variants (SNVs), and microsatellites, all of which have been used to reconstruct lineages (Woodworth et al., 2017). Somatic mutation in LINE-1 (L1) elements was used to track clonal lineages in human brain (Evrony et al., 2015). Somatic mutation in mitochondrial DNA have been shown to reconstruct cellular lineages with high sensitivity and specificity in human cells (Ludwig et al., 2019; Xu et al., 2019). While somatic mutations are endogenous thus can be extended to study human tissues directly, somatic mutations by nature occur at low frequency and therefore limit the reconstruction resolution to a clonal level.

With current limitations of existing lineage reconstruction methods, an endogenous lineage reconstruction method that yields cellular lineages at a single cell-division resolution is required for the study of tissue development and cell fate differentiation in humans.

C. *Cell characterization*

To study cell differentiation and tissue development, in addition to cellular lineages, information about the cells' characteristics or cell types also need to be identified. Cells of different cell types express different genes and have different transcriptomic profiles. Cell type identification based on single-cell RNA sequencing involves partitioning the cells into clusters. The cells in each cluster share a similar transcriptomic profile, making them distinct from the cells in other clusters. Generally, each cell cluster will contain one or more marker genes, which are genes that are expressed primarily in a single cell type. All cells in that cluster will be annotated as that one cell type. Because cell type annotation entirely depends on the types and quantities of RNA transcripts measured in each cell, a complete transcriptomic profile of each cell must be faithfully captured to allow accurate cell type identification.

1. Noncoding RNAs

Even though majority of transcriptomic studies focus on mRNA, as they code for proteins, there are many other types of RNA, collectively called noncoding RNA (ncRNA). In prokaryotic organisms, the majority of the genome is coding. On the other hand, only about 2–5% of mammalian genomes produce proteins, while about 90% is transcribed over the lifespan into many types of ncRNAs (Hubé and Francastel, 2018). The understanding of these transcripts' functions in eukaryotic cells is still incomplete.

There are two types of ncRNAs: housekeeping ncRNAs (tRNA and rRNA), which are present at higher level in cells and are directly involved in protein synthesis, and regulatory ncRNAs, which are further classified based on their size. Long ncRNA (lncRNA) have at least 200 nucleotides, while small ncRNA are shorter than 200 nucleotides. Small RNA are further divided into many subtypes. One of which is microRNA (miRNA), which can inhibit gene expression by binding to target mRNA. Many miRNAs play significant roles in cancer, where oncogenic miRNAs can regulate unique target genes, leading to tumorigenesis and tumor progression (Chatterjee and Wan). Another functional small RNA is PIWI-interacting (piRNA), which regulate transposon expression in the germ cells (Chatterjee and Wan). Therefore, in order to fully characterize cells within a tissue during development, ncRNA also need to be detected and analyzed along with mRNA. However, ncRNA are still understudied and there is a limited number of tools to measure and annotate them. Therefore, a comprehensive method to measure the entire transcriptomic profile of a cell holds a great promise to provide insights into the developmental process.

2. rRNA Depletion

In order to measure mRNA and ncRNA in a cell and capture its transcriptomic state, rRNA must first be depleted before sequencing. About 80% of total RNA in growing mammalian cells is estimated to be rRNA, 15% tRNA, and the rest is mRNA and other types of RNA (Lodish et al., 2000a). Due to the fact that a small portion of total RNA is mRNA, despite the decreasing cost of sequencing, it is still necessary to select for mRNA or select against tRNA and rRNA in a sample. The strategies to do so differ between eukaryotic cells (e.g. mammalian cells) and prokaryotic cells (e.g. bacterial cells).

Eukaryotic mRNA precursors are processed by 5' capping, 3' cleaving and polyadenylation, and RNA splicing to remove introns before being transported to the cytoplasm and translated (Lodish et al., 2000b). Therefore, all mRNA present in the cytoplasm have 3' polyadenylation, or a poly(A)-tail , which distinguishes them from other types of RNA. This poly(A)-tail allows mRNA to be selectively captured and enriched with oligo-dT primers (Hashimshony et al., 2016).

On the other hand, bacterial mRNA does not possess a poly(A)-tail and there is no distinct characteristic differentiating it from tRNA and rRNA. Fortunately, tRNA's stable secondary structure and utilization of many post-transcriptionally modified nucleotides naturally select against tRNA detection (Cozen et al., 2015; Zheng et al., 2015). As for rRNA, there are established strategies to deplete rRNA and enrich mRNA with varying success (Culviner et al., 2020; He et al., 2010; Huang et al., 2020; Kraus et al., 2019; Prezza et al., 2020). However, most strategies suffer from one or more of the following limitations: 1. They cannot be easily expanded to diverse bacterial species; 2. They need comparatively high starting total RNA material; 3. They require complex design; or 4. The depletion step's cost is relatively high. Therefore, an efficient and cost-effective bacterial mRNA sequencing tool that can be scaled down to lower starting materials is required to capture the transcriptomic profile of bacterial cells.

Because both bacterial mRNA and mammalian ncRNA lack poly(A)-tails, this bacterial mRNA sequencing tool can be extended to capture mammalian short ncRNA, while selecting against rRNA. With both mRNA and ncRNA information, a more complete transcriptomic profile for mammalian cells can be captured.

D. *In this dissertation*

To study cell differentiation and tissue development, both lineage relationship between cells within the system and cell type information are required. In Chapter 2, an endogenous lineage reconstruction technique that accurately predicts cellular relationship at a single cell-division resolution in early mouse embryos is presented. In Chapter 3, an efficient rRNA depletion protocol to enrich bacterial mRNA is discussed. In Chapter 4, we show that the method presented in Chapter 3 can be further optimized and be applied to mammalian cells to measure ncRNA, enabling a more complete transcriptomic profile for better cell characterization. Chapter 5 concludes the dissertation and discusses what work is required to further advance the field.

2. A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing

A. *Introduction*

Understanding lineage relationships between cells in a tissue is one of the central questions in biology. Reconstructing lineage trees is not only fundamental to understanding tissue development, homeostasis and repair but also important to gain insights into the dynamics of tumor evolution and other diseases. Genetically encoded fluorescent reporters have been a powerful approach to reconstruct the lineage of many tissues (Kretzschmar and Watt, 2012). However, these methods require the generation of complex animal models for each stem or progenitor cell type of interest, and are limited to a clonal resolution (Kretzschmar and Watt, 2012). Similarly, other pioneering techniques, such as the use of viruses (Naik et al., 2013), transposons (Sun et al., 2014; Wagner et al., 2018), Cre-loxP based recombination (Pei et al., 2017) and CRISPR-Cas9 (Alemany et al., 2018; Kalhor et al., 2017; McKenna et al., 2016; Perli et al., 2016; Raj et al., 2018; Spanjaard et al., 2018) have also been used to genetically label cells to primarily reconstruct clonal lineages that lack the resolution of an individual cell division. This clonal resolution limits our ability to understand tissue dynamics at a single cell-division resolution. While a recent report that combined CRISPR-Cas9-mediated targeted mutagenesis with single-molecule RNA fluorescence *in situ* hybridization enabled reconstruction of lineages at a single cell-division resolution (MEMOIR) (Frieda et al., 2017), the ability of the method to infer lineages dropped substantially by the 3rd cell division.

Further, as all these methods involve exogenous labeling strategies, they cannot be used to map cellular lineages in human tissues directly, thereby posing a significant barrier to understanding human development and diseases. While endogenous somatic mutations have been used to reconstruct lineages, the low frequency of their occurrence and distribution over the whole genome make them challenging to detect and therefore limit their application as a lineage reconstruction tool (Behjati et al., 2014; Ju et al., 2017; Lodato et al., 2015). Similarly, recent methods have used mutations within the mitochondrial genome or microsatellites to reconstruct lineages, but as most other lineage reconstruction approaches, it is limited to a clonal resolution (Biezuner et al., 2016; Evrony et al., 2015; Ludwig et al., 2019; Xu et al., 2019). Previously, we developed a method to detect the endogenous epigenetic mark 5-hydroxymethylcytosine (5hmC) in single cells (scAba-Seq) and showed that the lack of maintenance of this mark during replication coupled with the low rates of Tet-mediated hydroxymethylation resulted in older DNA strands containing higher levels of 5hmC (Mooijman et al., 2016). The ability to track individual DNA strands through cell division allowed us to deterministically reconstruct lineages that were limited to 2 cell divisions (Mooijman et al., 2016). Therefore, to reconstruct larger trees and to overcome limitations of other existing methods, we report scPECLR (single-cell Probabilistic Endogenous Cellular Lineage Reconstruction), a generalized probabilistic framework for endogenously reconstructing cellular lineages at an individual cell division resolution using single-cell 5hmC sequencing. We show that this approach can be used to successfully reconstruct up to 4 cell divisions. To reconstruct larger lineage trees from billions of possible tree topologies, we developed a new integrated single-cell method scH&G-seq to simultaneously sequence 5hmC and genomic/mitochondrial DNA from the same cell. By

combining information from genomic variants that can be used to identify clonal subtrees within the complete tree, together with strand-specific 5hmC that enables tracking the lineage of individual cells, scH&G-seq can be generalized to endogenously reconstruct the lineage of large trees at a single cell division resolution.

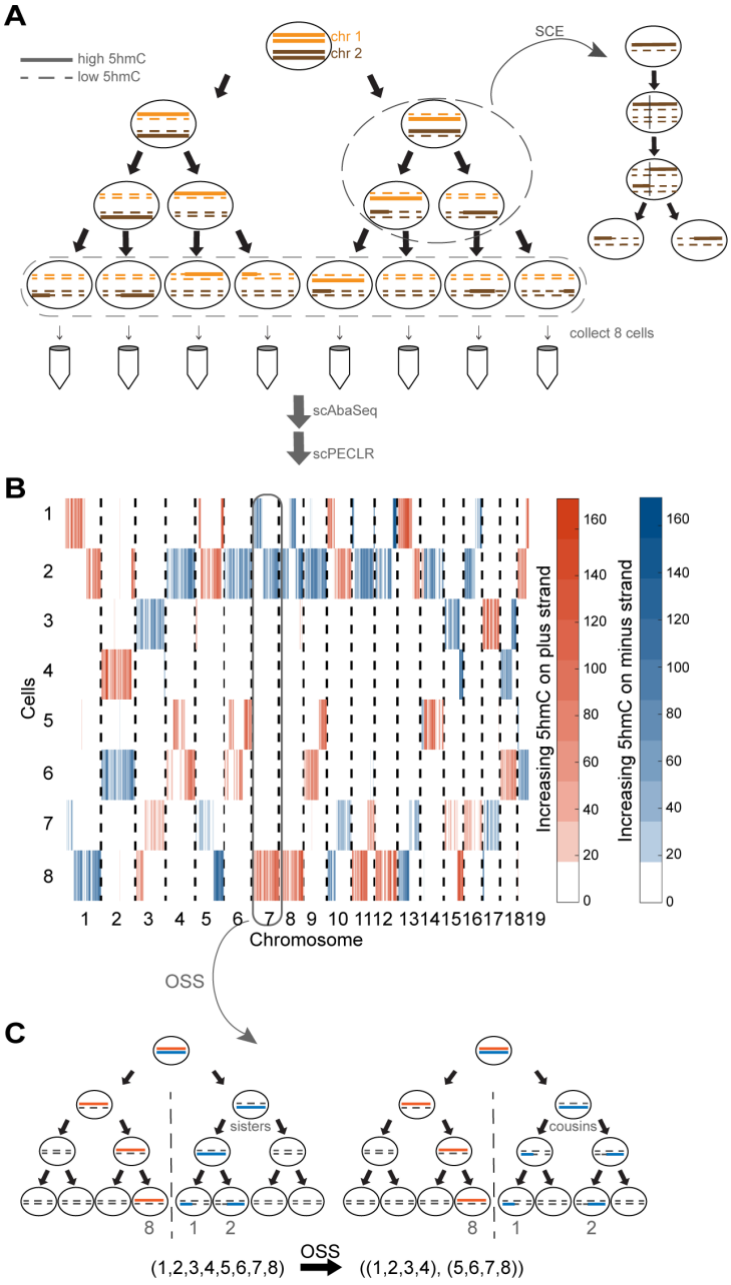


Figure 2.1 Strand-specific single-cell 5hmC data enables initial lineage bifurcation of individual cells into two subtrees

(A) Schematic shows a zygote with chromosomes containing high 5hmC levels (solid lines) undergoing three cell divisions. The newly synthesized strands at each cell division contain very low levels of 5hmC (dotted lines). SCE events occur randomly during each cell cycle. All cells are isolated and sequenced using scAba-Seq to quantify strand-specific 5hmC in single cells.

(B) Data shows mosaic pattern of strand-specific 5hmC in single cells obtained from an 8-cell mouse embryo. 5hmC counts within 2 Mb bins on the plus and minus strands of all chromosomes are shown in orange and blue, respectively.

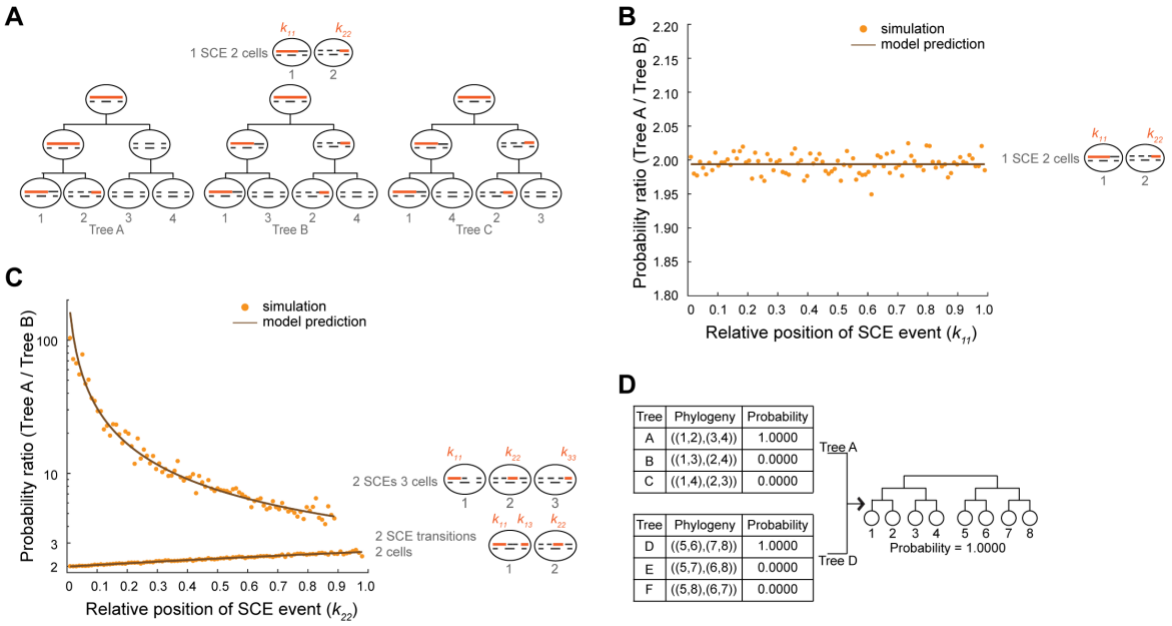
(C) The original paternal plus and minus strand of each chromosome should be found in cells on opposite sides of the lineage tree. This OSS analysis on chromosome 7 places cell 8 in one 4-cell subtree and cells 1 and 2 in the other subtree. Performing OSS on all chromosomes places cells in one of these two 4-cell subtrees and reduces the complexity of the lineage reconstruction problem.

B. *Genome-wide strand-specific 5hmC enables initial lineage bifurcation of individual cells into two subtrees*

As proof-of-principle, we dissociated 8-cell mouse embryos and performed scAba-Seq to quantify strand-specific genome-wide patterns of 5hmC in single cells (Figure 2.1A). As shown previously, a majority of 5hmC is present on the paternal genome during these stages of preimplantation development (Inoue and Zhang, 2011; Iqbal et al., 2011; Wossidlo et al., 2011). Single cells from an 8-cell embryo displayed a mosaic genome-wide distribution with no overlap of 5hmC between the plus and minus strands of a chromosome (Figure 2.1B). Further, we found that for each chromosome, the strand-specific 5hmC was localized to a few cells with other cells containing undetectable levels of the mark (Figure 2.1B). These observations clearly demonstrate that only one allele carries a majority of 5hmC, and that, consistent with previous results, we are primarily detecting 5hmC on the original paternal genome, with DNA strands synthesized in subsequent rounds of replication carrying very low levels of the mark. We used this as our basis to reconstruct cellular lineages of 8-cell mouse embryos.

As the first step towards reconstructing lineage trees, we noted that the original plus and minus strands of each paternal chromosome in the 1-cell zygote will be found in distinct cells on opposite sides of the lineage tree after *n* cell divisions. As a result, all cells can be

placed in one of two subtrees, thereby reducing the number of cell divisions to be reconstructed from *n* to *n-1*. For example, at the 8-cell stage, the original paternal plus strand of chromosome 7 is detected in cell 8 while the corresponding minus strand is detected in cells 1 and 2 (Figure 2.1B). This suggests that cell 8 is on the opposite side of the lineage tree compared to cells 1 and 2. Performing this first step of scPECLR, which we refer to as original strand segregation (OSS) analysis, over all the chromosomes enables us to systematically place cells 1-4 and 5-8 on opposite sides of the lineage tree for this embryo, reducing the complexity of the problem from reconstructing 3 cell divisions with 315 tree topologies to 2 cell divisions with 9 tree topologies (Figure 2.1C).



Tree	Phylogeny	Probability
A	((1,2),(3,4))	1.0000
B	((1,3),(2,4))	0.0000
C	((1,4),(2,3))	0.0000

Tree	Phylogeny	Probability
D	((5,6),(7,8))	1.0000
E	((5,7),(6,8))	0.0000
F	((5,8),(6,7))	0.0000

Figure 2.2 Endogenous 5hmC based lineage reconstruction using scPECLR

(A) Schematic showing that two cells sharing an original DNA strand (solid orange line) can either be sisters (Tree A) or cousins (Trees B and C) depending on whether the SCE event occurred at the 4- to 8- or 2- to 4-cell stage, respectively. All newly synthesized DNA strands are shown as dashed black lines.

(B) For the case of a SCE transition between two cells, the probability of the pair of cells being sisters *vs.* cousins is plotted against the relative position of the SCE event on the chromosome (k_{11}). The model prediction (black) and simulation results (yellow) are shown for chromosome 1 ($N = 97$ for 2 Mb bins) with $b = 0.3$.

(C) The probability ratio between Trees A and B are shown as a function of k_{22} for $N = 97$ and $b = 0.3$ for two cases: 2 SCE transitions shared between 2 cells and 2 SCE events shared between 3 cells.

(D) For the 8-cell mouse embryo shown in Figure 1B, the probability of observing the different topologies, rounded to four decimal places, of the two 4-cell subtrees are shown.

C. *Probabilistic lineage reconstruction using scPECLR accurately predicts 8-cell embryo trees*

To reconstruct the complete lineage tree, we next focused attention on the mosaic pattern of 5hmC arising from abrupt transitions in hydroxymethylation levels among cells along the length of a chromosome. As described previously, these sharp transitions in 5hmC that are shared between two cells are the result of homologous recombination during sister chromatid exchange (SCE) events in the G2 phase of a previous cell cycle (Mooijman et al., 2016). Detection of 5hmC transition points that are common to two cells therefore indicate a shared evolutionary history between these cells (Figure 2.1A, inset). However, while a SCE event at the 4-cell stage would imply that the cells are sister cells (Figure 2.1C, left), one occurring at the 2-cell stage would indicate that the same pattern of 5hmC transition can also be observed between cousin cells (Figure 2.1C, right). Thus, the observation of a single shared SCE event between two cells cannot be used to immediately discriminate between sister and cousin cell configurations.

To systematically determine the likelihood of observing different tree topologies, we developed a probabilistic framework where the occurrence of SCE events are modeled as a Poisson process. The total number of SCE events is used to estimate the parameter b of the Poisson process, the rate of SCE events per chromosome per cell division, using maximum likelihood estimation (MLE) (Appendix 0). Following OSS, 8-cell trees can be grouped into two 4-cell subtrees, each with 3 possible tree arrangements (Figure 2.2A). Next, we used the probabilistic model to calculate the likelihood of observing a SCE pattern for a chromosome given a tree topology. We observed a large variety of SCE patterns, ranging from commonly observed patterns, such as one or two SCE transitions shared between two cells, to more

complex distributions of 5hmC between cells (Appendix 5.D1: Figure S2.1). For the most common pattern of one SCE transition between two cells, scPECLR predicts that the tree with the two cells as sisters (Tree A) is twice as likely as one where the two cells are cousins (Tree B or C), in good agreement with simulated data (Figure 2.2B and Appendix 0). Similarly, both our model prediction and simulations show that when two SCE transitions are shared between two cells, the probability that the two cells are sisters is 2 to 3 times higher than the probability that they are cousins, with the likelihood ratio between sister and cousin tree configurations depending on the relative position of the SCE transition on the chromosome (Figure 2.2C and Appendix 0). The lower probability of observing this pattern in cousin tree arrangements arises from the constraint that only even number of SCE transitions can occur within the region between k_{11} and k_{13} of the chromosome during the last cell division (Appendix 5.D1: Figure S2.2A). More complex 5hmC distribution patterns, such as when two SCE events are shared between three cells substantially favors the Tree A configuration (Figure 2.2C and Appendix 0). After the SCE pattern of each chromosome is analyzed, we can estimate the total likelihood of observing different tree topologies, assuming that the SCE events on each chromosome are independent (Appendix 0). Finally, the likelihood of an 8-cell tree is the product of the likelihoods of the two corresponding 4-cell subtrees (Figure 2.2D). An automated pipeline to reconstruct cellular lineages is provided with this work (Appendix 0).

To test the accuracy of scPECLR, we simulated 5hmC patterns of 8-cell embryos with a SCE rate similar to the experimentally observed value ($b = 0.3$), which is also within the range of SCE event rates found in various other cell types (Falconer et al., 2012; Hongslo et al., 1991; Tateishi et al., 2003; Wu et al., 2017; Zack et al., 1977). We found that scPECLR

predicted the lineage tree correctly in 96% of all simulations (Figure 2.3A, left). In contrast, MEMOIR predicted the lineage tree accurately in only ~67% of the top 40% most reliably reconstructed trees, though this was based on ground truth obtained from imaging data (Figure 2.3A, left). This improved accuracy of scPECLR strongly suggests that endogenous strand-specific 5hmC patterns present an accurate tool to reconstruct lineage trees at an individual cell division resolution. Further, to directly validate our technique against experimental data, the lineage trees predicted by scPECLR from simulated 8-cell embryos were combined to estimate the number of SCE events at the 4-cell stage of development. We hypothesized that if scPECLR predicted the correct tree then it would produce a similar distribution of SCE events to experimental data at the 4-cell stage. Comparison of the scPECLR predicted distribution of SCE events per cell at the 4-cell stage was statistically not different from the experimentally obtained distribution in 4-cell embryos ($p > 0.8$, Two-sample Kolmogorov-Smirnov (KS) test) (Figure 2.3A, right). In contrast, when one of the 314 incorrect tree topologies at the 8-cell stage were sampled randomly, it resulted in a distribution of SCE events per cell that was statistically significant from the experimental data ($p < 10^{-4}$, Two-sample KS test) (Figure 2.3A, right). These results show that scPECLR can reconstruct 3 cell divisions with high accuracy. Finally, we applied scPECLR on the 8-cell mouse embryo shown in Figure 2.1B and other embryos to predict lineage trees with high confidence (Figure 2.2D and Appendix 5.D1: Figure S2B).

As SCE transitions play a central role in reconstructing cellular lineage trees with scPECLR, we next explored how the endogenous rate of SCE events influences the accuracy of the model. As expected, the accuracy of lineage reconstruction increases monotonically with increasing rates of SCE events, with greater than 98% of the simulated 8-cell trees

correctly predicted for $b \geq 0.4$ (Figure 2.3B and Appendix 0). These simulations were performed using 19 paternal autosomes, consistent with our observation that a majority of 5hmC is found on the paternal genome in preimplantation mouse embryos. However, most cell types carry 5hmC on both parental genomes and therefore, we also performed simulations with 38 chromosomes. Again, as expected, the predictive power of the model increases, with more than 98% of the simulated 8-cell trees accurately predicted for $b \geq 0.2$ (Figure 2.3B). These results demonstrate that the lineage tree can be accurately predicted up to 3 cell divisions even with low rates of SCE events (Figure 2.3B).

Figure 2.3 scPECLR can reconstruct 8-cell lineage trees accurately and be extended to reconstruct larger lineage trees

(A) (*Left*) scPECLR accurately predicts the lineage of 96% of simulated 8-cell trees ($b = 0.3$). Error bars indicate the bootstrapped standard error. In comparison, MEMOIR accurately predicts 67% of the top 40% most reliably reconstructed 8-cell trees (Frieda et al., 2017). (*Right*) The distribution of SCE events in 4-cell embryos (blue) is not statistically different from that of 4-cell trees inferred with scPECLR starting from 8-cell embryos (orange, $p > 0.8$), but is different from that of 4-cell trees inferred starting from a random topology at the 8-cell stage (brown, $p < 10^{-4}$).

(B) Panel shows the percentage of simulated 8-cell and 16-cell trees that are correctly predicted by scPECLR for different SCE rates (b). Solid and dotted lines indicate cells where 5hmC can be quantified in 19 or 38 chromosomes, respectively. The prediction accuracy is computed by simulating 5000 trees. Error bars indicate the bootstrapped standard error of prediction accuracy.

(C) Panel shows the percentage of 2-, 4- and 8-cell subtrees that are accurately predicted within simulated 16-cell trees as a function of the SCE rate (b). The prediction accuracy is computed by simulating 5000 16-cell trees. Error bars indicate the bootstrapped standard error of prediction accuracy.

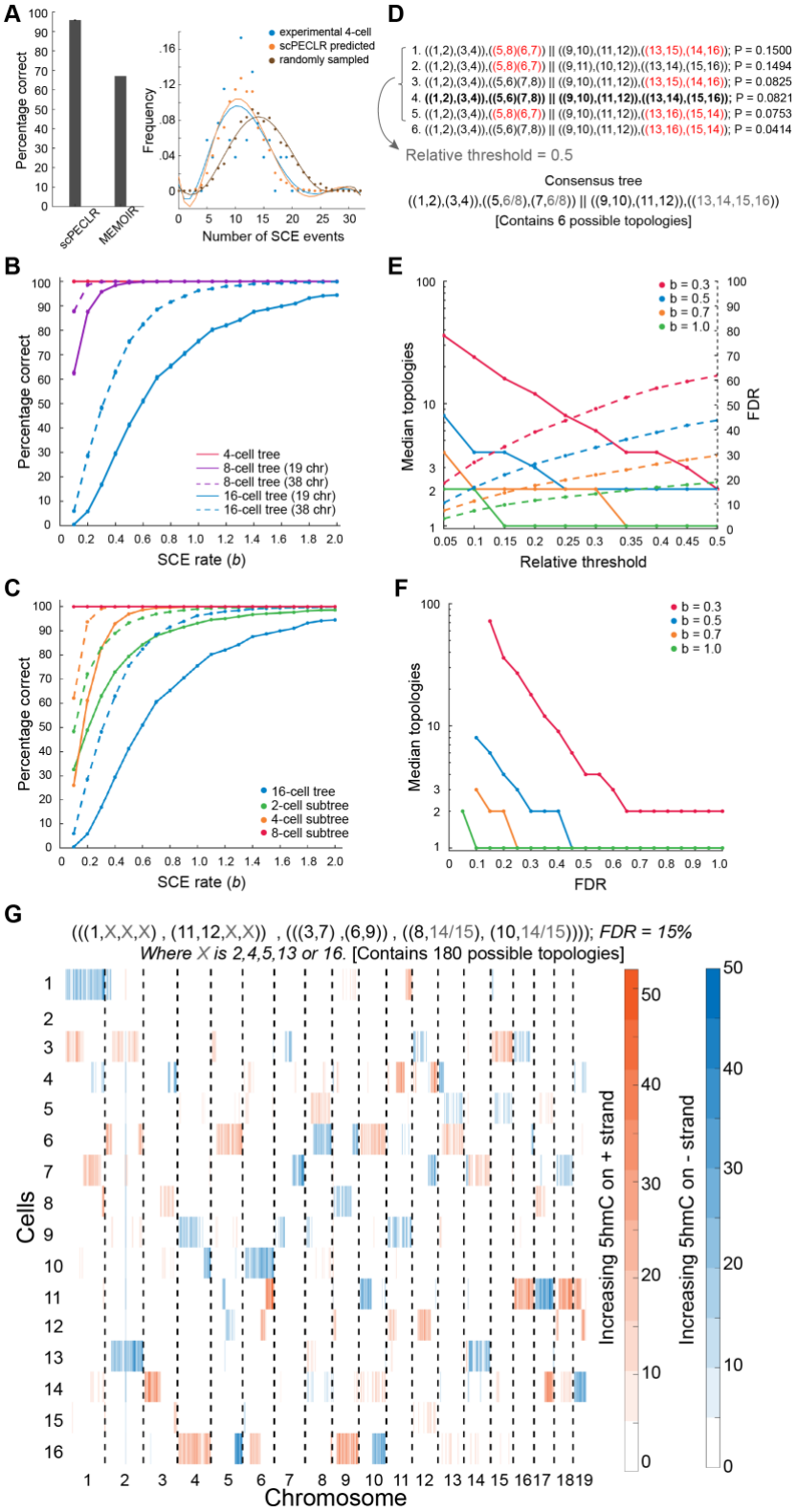
(D) Schematic illustrating how consensus trees are obtained. In this example, the top 6 tree topologies (with the highest probabilities) that are obtained after applying scPECLR on a simulated 16-cell tree are shown. The relative threshold parameter is used to determine the number of topologies that are considered in the consensus tree analysis. With a relative threshold of 0.5, the top 5 tree topologies in this example are selected to generate a consensus tree that is consistent with all these trees. The uncertainty within the consensus tree is quantified by the number of tree topologies it contains. The higher the number of tree topologies it contains, the higher is the uncertainty within the consensus tree. Red fonts indicate parts of the lineage tree that are incorrectly predicted. The tree highlighted in bold is the true tree.

(E) Simulation results show that as the relative threshold increases, the median number of topologies in the consensus tree decreases (solid lines, left axis), while the false discovery rate (FDR) increases (dotted lines, right axis). For these simulations, two other parameters $t8$ and $t4$ are set to 0.75 and 1.0, respectively. (For additional details on the parameters, see Appendix 0).

(F) As the FDR decreases, the median number of topologies contained within the consensus tree increases. Thus, this panel shows how the specificity of the consensus tree is related to error tolerance. For $b \geq 0.7$, the median number of topologies contained within the consensus tree rapidly drops to 1, suggesting that the consensus tree is fully constrained and is the correct tree. Note, the lowest FDR possible for b = 0.3, 0.5, 0.7 and 1.0 are 15%, 10%, 10%, and 5%, respectively.

(G) Single-cell 5hmC sequencing data is shown for a 16-cell mouse embryo (4 Mb bins). The consensus tree associated with this embryo is estimated to have a 15% FDR rate. Relative threshold, $t8$, and $t4$ are set at 0.05, 0.85 and 0.8, respectively (for additional details on the parameters, see Appendix 0). The consensus tree

is constrained to only 180 possible topologies, a significant reduction from the more than 600 million trees originally.



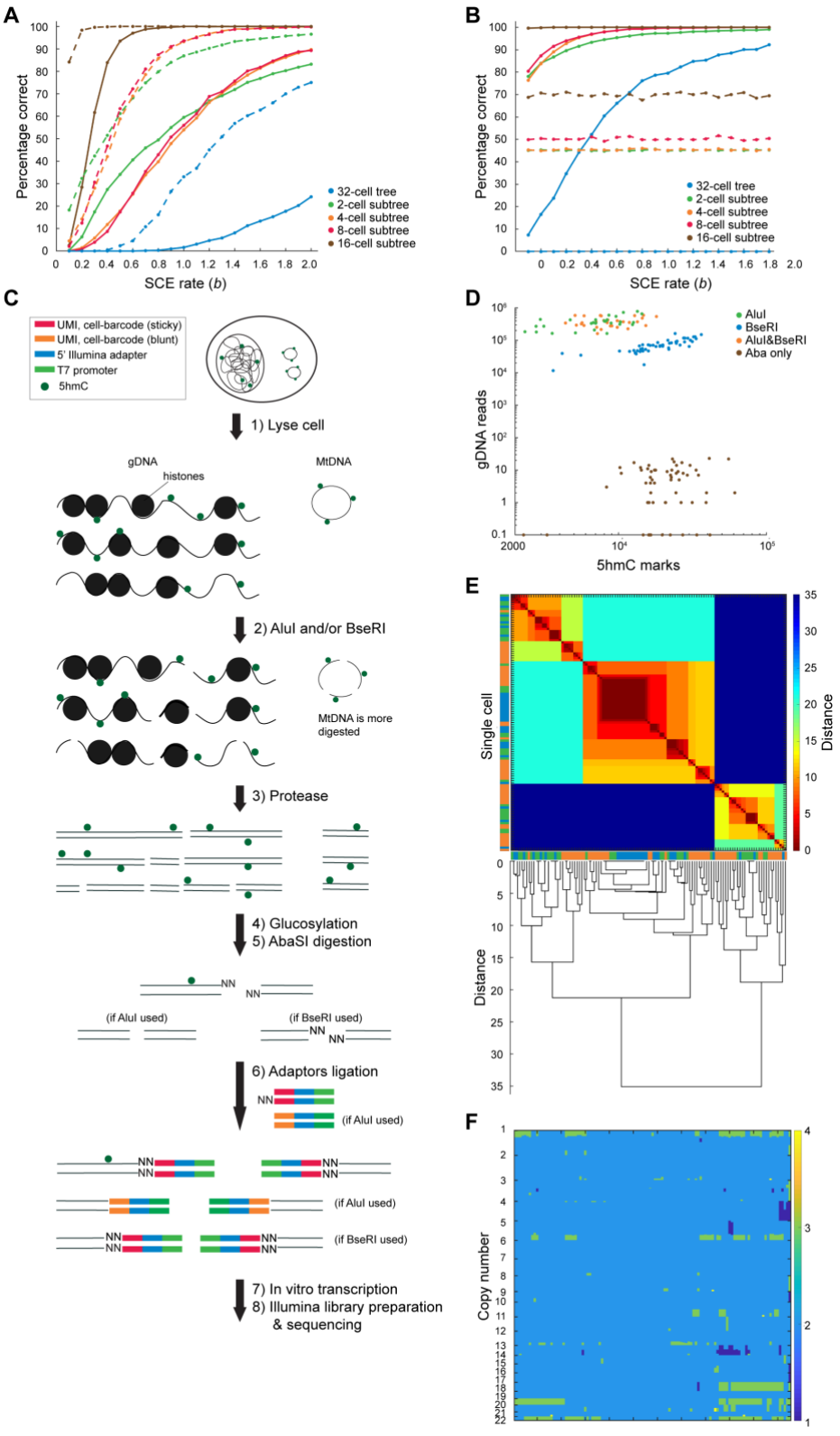


Figure 2.4 Integrated single-cell 5hmC and genomic DNA sequencing can be used to endogenously reconstruct larger lineage trees at an individual cell division resolution with improved accuracy

(A) Panel shows the percentage of the full lineage, along with 2-, 4-, 8-, and 16-cell subtrees, that are accurately predicted within simulated 32-cell trees as a function of SCE rates (b). The prediction accuracy is computed by simulating 2000 trees. Solid and dotted lines indicate cells where 5hmC can be quantified in 19 or 38 chromosomes, respectively.

(B) Panel shows the percentage of the full lineage, along with its subtrees, that are correctly predicted in simulated 32-cell trees as a function of SCE rates (b), using information from both 5hmC and gDNA. Solid lines indicate the prediction accuracy using integrated information, while the dotted lines indicate the accuracy using gDNA alone. The prediction accuracy is computed by simulating 2000 38-chr trees, and the rate of occurrence of genomic variants is set to 0.6 per chromosome per cell division.

(C) Schematic illustrating scH&G-seq. Restriction enzyme(s) are used to digest both gDNA and mtDNA prior to the protease step to enrich for mtDNA. These steps are then followed by glucosylation, digestion with AbaSI, ligation of double-stranded adapters, IVT amplification and Illumina library preparation.

(D) scH&G-seq using AluI, BseRI, or both enzymes enable detection of genomic DNA while simultaneously capturing 5hmC sites in the genome.

(E&F) Single cells from three sequencing libraries cluster into two major groups. (E) Panel shows the heatmap of the Euclidean distance between the cells and the corresponding dendrogram. Cells from AluI, BseRI, and dual enzyme libraries are displayed in green, orange, and blue, respectively. (F) Panel show the heatmap of the copy number profile of single cells sorted in the same order as the dendrogram shown above.

D. *scPECLR can be extended to reconstruct the lineage of 16-cell trees*

We next extended scPECLR to reconstruct the lineage of 16-cell trees, where the number of possible tree topologies increase exponentially to more than $6x10^8$. While the ability to predict the complete lineage tree decreases (17% of all simulated 16-cell trees were predicted correctly for $b = 0.3$), we found that in a majority of cases large parts of the lineage tree were reconstructed accurately with the most common error being the misidentification of one sister pair within a 4-cell subtree (Figures 2.3B and 2.3C). For a SCE rate of $b = 0.3$, 83% of all 4-cell subtrees and 63% of all 2-cell subtrees (sister pairs) were predicted correctly (Figure 2.3C). These results suggest that when reconstructing 16-cell trees from strand-specific 5hmC data, it will be important to identify parts of the lineage tree that we can predict with high confidence. To accomplish this, we first included all tree topologies that were predicted to have probabilities above a threshold relative to the tree with the highest probability (Figure 2.3D). A consensus tree that is consistent with all these tree topologies is then established (Figure 2.3D, Appendix 5.D1: Figure S2.3 and Appendix 0). As the relative threshold is increased (that is, we include fewer tree topologies to construct the consensus tree), the median

consensus tree contains fewer topologies, resulting in a more specific or constrained consensus tree. However, this comes at the expense of an increase in false discovery rate (FDR). For example, with $b = 0.3$ and for a relative threshold of 0.1, the median consensus tree contained 24 tree topologies (Figure 2.3E, solid red line). The consensus trees displayed a FDR of ~26%, implying that in 26% of the simulations, the consensus tree has at least some part of the lineage tree that was not consistent with the true tree (Figure 2.3E, dotted red line). Thus, the relative threshold allows us to tune the competing goals of specificity and accuracy of the consensus tree. These results show that for a certain rate of SCE events and a desired level of FDR, the median number of topologies contained in the consensus tree can be estimated, yielding insights into how much lineage information can be extracted from the 5hmC data based on the number of SCE events and our error tolerance (Figure 2.3F and Appendix 0). Finally, as proof-of-principle, we sequenced a 16-cell mouse embryo and applied scPECLR to show that we can extract partial lineage information from larger trees (Figure 2.3G and Appendix 0).

E. *Integrated single-cell genomic DNA and 5hmC sequencing enables reconstruction of larger lineage trees*

For larger 32-cell trees, the number of possible tree topologies increase to more than 10^{26}, making it computationally very expensive to calculate the likelihood of all possible trees. Therefore, we next extended scPECLR by developing an algorithm that efficiently searches through the tree topology space to reconstruct these larger lineage trees. After OSS bifurcates the 32 cells into 2 16-cell subtrees, we identify groups of 8 cells that when combined minimize the number of SCE events at the 4-cell stage. This algorithm relies on the strategy that incorrectly grouped cells will increase the number of SCE events at the 4-

cell stage, and this sub-sampling enables rapid search through the tree topology space. Finally, the 4 groups of 8 cells are reconstructed using scPECLR as described in the previous sections (Appendix 0). As expected, while the ability to predict the complete lineage tree is lower than that for 16-cell trees, this method is able to rapidly predict sub-trees within the 32-cell tree. For example, for $b = 1$ and 19 alleles, 2-, 4- and 8-cell subtrees are predicted with 50-60% accuracy while the 16-cell subtrees are predicted at close to 100% accuracy (Figure 2.4A, solid lines). For the more general case of 38 alleles in mouse genomes, the prediction accuracy increases substantially with 80-95% of the 2-, 4- and 8-cell subtrees predicted correctly for $b = 1$ (Figure 2.4A, dotted lines).

To endogenously reconstruct large lineage trees at an individual cell division resolution, we hypothesized that single-cell strand-specific 5hmC data combined with information on genomic variants, such as genomic copy number variations (CNV), genomic single-nucleotide polymorphisms (SNP) or mitochondrial SNPs, could significantly improve the prediction accuracy. Genomic variants have previously been used to reconstruct clonal lineages and therefore, when integrated with strand-specific 5hmC could help anchor subtrees within the complete lineage tree with high confidence (Biezuner et al., 2016; Evrony et al., 2015; Ludwig et al., 2019; Xu et al., 2019). To test this hypothesis, we simulated trees with genomic variants together with SCE events and found that the prediction accuracy increases dramatically compared to the use of SCE events alone in predicting trees (Figures 2.4A, 2.4B, S2.4B and S2.4C) (Appendix 0). For example, for $b = 1$, the complete 32-cell lineage tree was predicted correctly in 76% of all the simulations and the 2- to 16-cell subtrees were predicted with greater than 96% accuracy (Figure 2.4B). In contrast, when using strand-specific 5hmC or genomic variants alone, the prediction

accuracy was lower (Figures 2.4A,B). Overall, these results demonstrate that integrated data from 5hmC and genomic variants present a general strategy to accurately reconstruct large lineage trees at a single cell division resolution.

To accomplish this goal experimentally, we developed a new integrated technology to simultaneously quantify 5hmC and genomic DNA from the same cell (scH&G-seq). Single cells, sorted into 384-well plates using FACS are lysed, and the genomic DNA and mitochondrial DNA are digested using the restriction enzymes AluI and/or BseRI (Figure 2.4C). Our strategy for identifying these restriction enzymes that were optimal and compatible with this new technology is provided in Appendix 0. Next, after stripping chromatin from genomic DNA, 5hmC sites in the genome are glucosylated using T4 phage β-glucosyltransferase, and these glucosylated sites are thereafter digested by the restriction enzyme AbaSI (Figure 2.4C). Double-stranded adapters, containing a cell-specific barcode, a 5' Illumina adapter and T7 promoter, together with restriction enzyme-compatible overhangs are ligated to the fragmented genomic DNA molecules (Figure 2.4C). These ligated molecules are then amplified by in vitro transcription and used to prepare Illumina libraries as described previously (Hashimshony et al., 2016; Mooijman et al., 2016; Rooijers et al., 2019). Information on the restriction enzymes used in the experiment together with the cell-specific barcodes are used to simultaneously quantify both genomic DNA/mitochondrial DNA and 5hmC from the same cell.

As proof-of-concept, we applied scH&G-seq to single H9 human embryonic stem cells with different combination of restriction enzymes – AluI and AbaSI, BseRI and AbaSI, or AluI, BseRI and AbaSI – and successfully detected both genomic DNA/mitochondrial DNA and 5hmC from the same cell (Figure 2.4D and S2.4D). When compared to the scAba-seq

control cells, we detected a similar number of 5hmC sites per cell, and integration with additional restriction enzymes enabled genome-wide sequencing of genomic and mitochondrial DNA (Figure 2.4D and S2.4D). To demonstrate that genomic DNA sequencing could be used to infer clonal cellular relationships, we used the circular binary segmentation algorithm to call CNVs in single cells. Hierarchical clustering identified 2 major clusters with a diploid and two non-diploid population, with additional subgroups within the non-diploid population (Figure 2.4E and 2.4F). Consistent with previous work in tumor and other cells types that shows CNVs can be used to identify clonal populations, this result suggests that integrated genomic DNA and 5hmC sequencing could be used to predict large lineage trees at a single cell division resolution. Similarly, the high mutation rate of mitochondrial DNA has previously been used to reconstruct clonal lineage trees, and therefore we used scH&G-seq to identify mitochondrial SNPs. While we identified almost 40 mitochondrial SNPs in H9 cells when mapping to the reference human genome, these SNPs were all observed at a frequency of close to 100%. Comparison to previously published ATAC-seq data from H9 cells together with SNP calls from another human cell line also identified the same SNPs, suggesting that these nucleotides represented the wild-type sequence (Appendix 5.E1: Supplementary Table S1.1) (Diroma et al., 2020; Liu et al., 2017). Nevertheless, these results provide proof-of-principle that in addition to sequencing 5hmC in single cells, scH&G-seq can be used to obtain clonal lineage information that can together be used to reconstruct larger trees.

F. *scPECLR can be used to infer the rate of SCE events at each cell division and test the "immortal strand" hypothesis*

In addition to reconstructing cellular lineage trees, scPECLR can also be used to infer the rate of SCE events at individual cell divisions. For example, for 8-cell mouse embryos, the 5hmC distribution at the 4-cell and 2-cell stages can be reconstituted based on the predicted lineage, enabling us to estimate the rate of SCE events at each cell division (Figure 2.2D and Appendix 5.D**Error! Reference source not found.**: Figure S2.2B). While the overall SCE rate over three cell divisions for all the 8-cell mouse embryos analyzed in this study was estimated to be 0.35 events per chromosome per cell division on average, the individual SCE rates for the 1-to-2, 2-to-4, and 4-to-8 cell stages were 0.31, 0.24, and 0.51, respectively. Further, we found that the different rates of SCE events at each cell division did not affect the prediction accuracy of scPECLR (Appendix 5.D1: Figure S2.5 and Appendix 0). These results show that scPECLR can be used to infer the rate of double-stranded DNA breaks at each cell division and that the rate of SCE events can vary during development.

Finally, we explored another application of scPECLR. As scPECLR uses endogenous strand-specific 5hmC in single cells to reconstruct 8-cell trees with high accuracy, we hypothesized that this method could be used to quantify how paternal alleles are segregated during cell division (Figure 2.5A). Different stem cell populations, such as hair follicle (Huh et al., 2013), neural (Karpowicz et al., 2005), satellite muscle (Conboy et al., 2007; Rocheteau et al., 2012) and intestinal crypt stem cells (Falconer et al., 2010; Potten et al., 2002), have previously been shown to display non-random segregation of DNA strands that can influence cell fate decisions. These results have led to the "immortal strand"

hypothesis that postulates old DNA strands are retained by daughter stem cells during asymmetric cell divisions to reduce the mutational load arising from genome replication of these longer lived cells. During mouse preimplantation development, recent reports have shown that blastomeres show biases in cell fate specification as early as the 4-cell stage (Goolam et al., 2016; White et al., 2016). Therefore, as proof-of-concept, we investigated sister chromatid segregation patterns of the paternal alleles at the 4-cell stage. To do this, we first combined 5hmC data from reconstructed sister cell pairs at the 8-cell stage to generate the distribution of the oldest DNA strands at the 4-cell stage (Figure 2.5B). In the example shown, when comparing cells (1,2) and (3,4), the original DNA strands appear to preferentially segregate to cell (1,2). In contrast, such a non-random pattern of DNA strand segregation is not observed between sister cells (5,6) and (7,8). Quantitatively, we analyzed 14 8-cell mouse embryos (equivalent to 28 2-to-4 cell division events) to find one sister pair at the 4-cell stage that displayed statistically significant non-random segregation of DNA strands ($p < 0.05$) (Figure 2.5C and Appendix 0). To directly validate these results, we also performed scAba-seq on single cells isolated from 13 4-cell mouse embryos (equivalent to 26 2-to-4 cell division events). We again observed a similar distribution of one sister pair that displayed a statistically significant non-random segregation pattern of DNA strands ($p < 0.05$), which was not significantly different from that observed in 8-cell embryos ($p > 0.8$, Two-sample KS test) (Figure 2.5C and Appendix 0). The observation of 2 non-random segregation events out of 27 embryos was not statistically significant ($p > 0.15$), suggesting that this level of non-random segregation at the 4-cell stage of mouse embryogenesis could arise by random chance (Figure 2.5D and Appendix 0). While more data points are required to validate non-random segregation in early mouse embryos, this proof-of-concept study

shows that strand-specific reconstruction of lineage trees can be a powerful approach to test the immortal strand hypothesis in different stem cell populations.

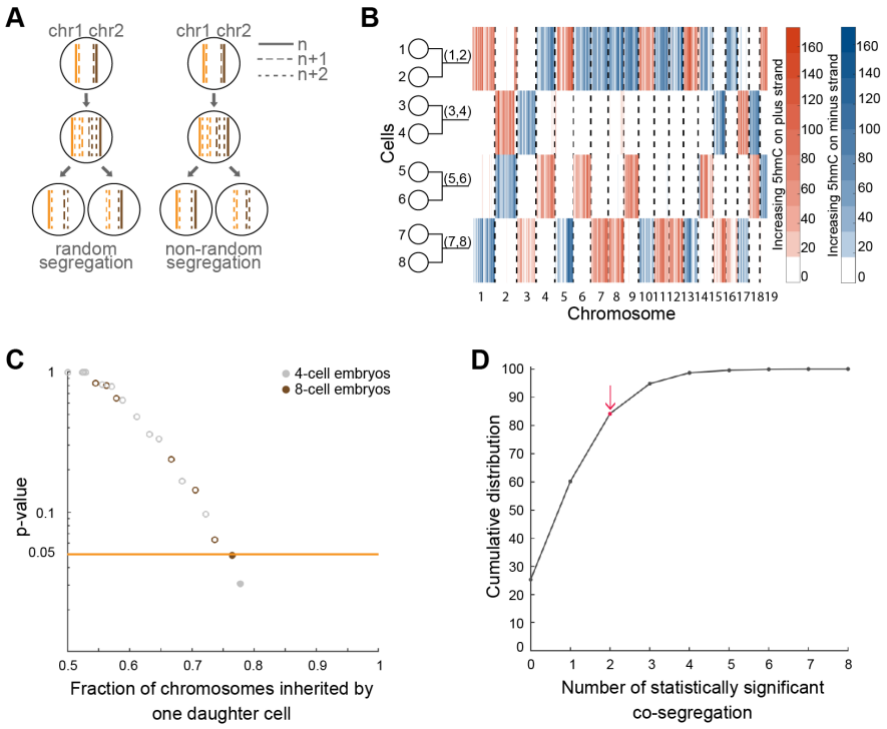


Figure 2.5 scPECLR can be used to map DNA strand segregation patterns

(A) Schematic showing DNA strand segregation patterns during cell division. Non-random segregation results in old DNA strands being preferentially inherited by one daughter cell. The oldest DNA strands are shown as solid lines, and strands synthesized in the $n+1$ and $n+2$ generation are shown as dashed and dotted lines, respectively.

(B) Combining the experimental 5hmC data for the 8-cell embryo in Figure 1B with the lineage tree predicted by scPECLR enables the genome-wide reconstitution of 5hmC in single cells at the 4-cell stage.

(C) Proof-of-principle testing non-random segregation of DNA strands at the 4-cell stage of mouse embryogenesis. The p-values from a binomial test under a null hypothesis of random segregation shows that, out of 27 embryos, two pairs of sister cells display statistically significant ($p < 0.05$) non-random segregation of DNA strands. Data obtained from 4- and 8-cell embryos are presented in gray and brown, respectively.

(D) 27 embryos were randomly sampled 10000 times from a pool of 100,000 simulated 4-cell embryos, generated with a constant SCE rate of $b = 0.3$. A cumulative distribution of the number of sister pairs that display statistically significant ($p < 0.05$) non-random segregation within the 27 embryos is shown. Red dot indicates the experimentally observed value of 2.

G. *Discussion*

Cellular lineage reconstruction plays an important role in answering fundamental questions in several areas of biology, such as immunology, cancer biology, and

developmental and stem cell biology. However, most current methods have two major limitations: (1) Clonal lineage reconstruction that cannot establish lineage relationships at the resolution of individual cell divisions; and (2) The use of transgenes that involves time-intensive generation of complex animal models and is an approach that cannot be extended to map lineages in human tissues. To overcome these limitations, we have developed a generalized probabilistic framework scPECLR to reconstruct short-term cellular lineage trees at an individual cell division resolution using strand-specific single-cell 5hmC sequencing data. Further, scPECLR can potentially also be combined with single-cell measurements of other non-maintained epigenetic marks, such as 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), to reconstruct lineages (Wu et al., 2017). In addition, the method could be extended to systems where the marks are not maintained through cell division, that is, the chromosome strands present in the original cell can be distinguished from subsequently synthesized strands, such as in systems exposed to bromodeoxyuridine (BrdU) (Claussin et al., 2017; Sanders et al., 2020). Finally, we show that by integrating 5hmC sequencing with information on genomic variants from the same cell (scH&G-seq) significantly improves the prediction accuracy of larger lineage trees. Most importantly, the use of an endogenous epigenetic mark and genomic variants to reconstruct lineage trees suggests that this method can be directly extended to study human development.

While scPECLR enables endogenous lineage reconstruction at a single cell division resolution, the method suffers from two limitations. First, it cannot be applied to cell types where the levels of 5hmC are below the detection limit of scAba-seq and scH&G-seq. However, as scPECLR relies on the relative levels of 5hmC between the two strands of a chromosome, it can be applied to many cell types, including those with low levels of 5hmC

in their genome provided that the kinetics of Tet-mediated hydroxymethylation are slow compared to one cell division time.

For example, this is evident in 16-cell stage mouse embryos that display distinct mosaic genome-wide strand-specific 5hmC patterns that enable lineage reconstruction despite undergoing global erasure of DNA methylation (Messerschmidt et al., 2014; Saitou et al., 2012) (Figure 2.3G). A second general limitation of reconstructing larger lineage trees at a single cell division resolution is that the number of tree topologies increase exponentially, resulting in a drop in prediction accuracy with each additional cell division. However, as this work demonstrates, scPECLR, in combination with scH&G-seq, significantly improves the lineage reconstruction accuracy of larger trees (Figure 2.4B). Finally, as most other lineage reconstruction methods resolve larger scale clonal information, scPECLR presents a complementary approach to these methods for applications that require reconstructing smaller lineage trees at an individual cell division resolution.

In the future, combining detection of 5hmC with measurements of mRNA from the same cell can potentially be used to simultaneously quantify both the cell type and the lineage relationship between cells in a tissue, thereby enabling us to directly probe symmetric and/or asymmetric cell fate decisions of stem cells at an individual cell division resolution. Additionally, combining CRISPR-Cas9 mediated genetic barcoding with scH&G-seq and mRNA-sequencing could enable large scale tracking of the dynamics of tissue development (Alemany et al., 2018; McKenna et al., 2016; Raj et al., 2018; Spanjaard et al., 2018). Overall, such measurements will provide detailed insights into how stem cells maintain an exquisite balance between self-renewal and differentiation to regulate the dynamics of tissue development and homeostasis. Finally, we anticipate that integrating 5hmC based lineage

reconstruction with measurements of other epigenetic marks from the same cells holds tremendous promise in understanding the genome-wide transmission and inheritance of the epigenome at each cell division.

3. Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion

A. Introduction

Bacterial species pervade our biosphere and millions of years of evolution have optimized these microbes to perform specific biochemical reactions and functions; processes that could potentially be adapted to develop a variety of products, such as renewable biofuels, antibiotics, and other value-added chemicals (Barajas et al., 2017; Dvořák et al., 2017; Kung et al., 2012; Otero and Nielsen, 2010; Peng et al., 2016). Bacterial messenger RNA (mRNA) sequencing provides a snapshot of the genome-wide state of a microbial population, and therefore enables fundamental understanding of these varied microbial functions and phenotypes (Creecy and Conway, 2015).

However, compared to eukaryotes, mRNA sequencing from bacterial samples has been more challenging for several reasons. First, unlike in eukaryotes, bacterial mRNA does not contain a poly(A)-tail at the 3' end that can be used to easily enrich for these molecules during reverse transcription (Mortazavi et al., 2008; Proudfoot, 2011). Further, total RNA isolated from bacterial cells typically contains greater than 95% ribosomal RNA (rRNA), and therefore cost-effective and high coverage sequencing of the transcriptome requires the development of efficient strategies to deplete the abundant 5S, 16S and 23S rRNA molecules (Giannoukos et al., 2012). Finally, bacterial cells typically contain approximately 100-fold lower RNA than mammalian cells, and as the starting amount of total RNA when working with rare, non-cultivable, and non-model bacterial species can be limiting, it is a

challenge to robustly and accurately capture the transcriptome from small quantities of total RNA with minimal amplification biases (Kang et al., 2011).

Several commercial kits have been developed to deplete bacterial rRNA from total RNA samples, including the MICROBExpress Bacterial mRNA Enrichment Kit (Thermo Fisher Scientific), the RiboMinus Transcriptome Isolation Kit, bacteria (Thermo Fisher Scientific), and the Ribo-Zero rRNA Depletion Kit (Illumina) (Petrova et al., 2017). These techniques rely on subtractive hybridization to deplete rRNA and typically work at a scale of hundreds of nanograms to micrograms of starting total RNA. Further, as these commercial kits are only effective on species targeted in the standard probe set, it is challenging to extrapolate these methods to diverse bacterial species (Giannoukos et al., 2012; Petrova et al., 2017). While this limitation of pre-designed kits have been overcome through the development of workflows to generate custom subtractive hybridization probe sets for any species of interest, they still operate at microgram quantities of starting material and either require multiple rounds of hybridization or a series of oligo optimization steps prior to optimal performance (Culviner et al., 2020; Kraus et al., 2019). An alternate approach relies on the Terminator™ 5’-phosphate-dependent exonuclease (TEX) (Lucigen) to specifically degrade rRNAs with 5’-monophosphate ends but not mRNAs with 5’-triphosphate ends; however, this method typically has lower efficiencies than other existing rRNA depletion strategies (He et al., 2010; Kang et al., 2011; Kuchina et al., 2019). A more recent method uses complementary single-stranded DNA probes to tile rRNAs that are subsequently degraded by RNase H (Huang et al., 2020). The commercial NEBNext Bacteria rRNA depletion kit (NEB) employs a similar strategy and can be applied to as low as 10 ng of starting total RNA. Similarly, another approach uses a pool of tiled single-guide RNAs to direct Cas9

mediated cleavage of rRNA-derived cDNA to deplete rRNA while another approach uses targeted reverse transcription primers designed to avoid capturing rRNAs (Armour et al., 2009; Prezza et al., 2020). However, all these methods require a large array of probes that can be expensive to synthesize and potentially need to be redesigned for distant bacterial species (Armour et al., 2009; Huang et al., 2020; Prezza et al., 2020).

Therefore, in this work we have developed EMBR-seq (Enrichment of mRNA by Blocked rRNA), a new technology that overcomes the limitations of sequencing mRNA from bacterial samples by: (1) Using 5S, 16S and 23S rRNA blocking primers and poly(A)-tail ing to specifically deplete rRNA and enrich mRNA during downstream amplification; (2) Using a single or a few blocking primer for each of the three abundant rRNA molecules, thereby enabling rapid adaptation to different bacterial species and significantly reducing the cost per sample; and (3) Using a linear amplification strategy to amplify mRNA from as low as 20 picograms of total RNA with minimal amplification biases. We applied EMBR-seq to a model *E. coli* system to demonstrate efficient mRNA enrichment and sequencing with increased sensitivity in gene detection. Further, we show that our method accurately captures the genome-wide gene expression profiles with minimal technical biases. Thus, EMBR-seq is an efficient and cost-effective approach to sequence mRNA from low-input bacterial samples.

B. Results

1. EMBR-seq uses blocking primers to deplete rRNA

To overcome the limitations described above, we developed EMBR-seq, a new technique to efficiently deplete rRNA from total RNA, thereby enabling cost-effective

sequencing of mRNA from bacterial cells. To minimize rRNA-derived molecules in the final sequencing library, we first incubated the total RNA with rRNA blocking primers, designed specifically to bind the 3' end of 5S, 16S and 23S rRNA, followed by poly-adenylation with *E. coli* poly-A polymerase (Figure 3.1 and Appendix 5.B). To deplete rRNA, EMBR-seq only requires primers at the 3' end of rRNA, unlike recent methods that tile oligonucleotides along the entire length of rRNA molecules, thereby significantly reducing costs and making our approach more easily translatable to other bacterial species. The blocking primers generate double-stranded RNA-DNA hybrid molecules at the 3' end of rRNAs, which reduces subsequent poly-adenylation and downstream amplification of rRNA molecules, as the poly-A polymerase preferentially adds adenines to single-stranded RNA (Feng and Cohen, 2000). Thereafter, the reaction mixture is reverse transcribed following the addition of a poly-T primer. This primer has an overhang containing a sample-specific barcode to enable rapid multiplexing and reduction in library preparation costs, the 5' Illumina adapter, and a T7 promoter (Hashimshony et al., 2016). After second strand synthesis, cDNA molecules are amplified by *in vitro* transcription (IVT). However, as only cDNA molecules deriving from a poly-adenylated RNA have a T7 promoter, our technique further amplifies mRNA-derived molecules for sequencing whereas rRNA-derived molecules are excluded from IVT amplification. The amplified RNA from IVT is then used to prepare Illumina sequencing libraries, as described previously (Figure 3.1 and Appendix 5.B) (Hashimshony et al., 2016; Mooijman et al., 2016; Rooijers et al., 2019).

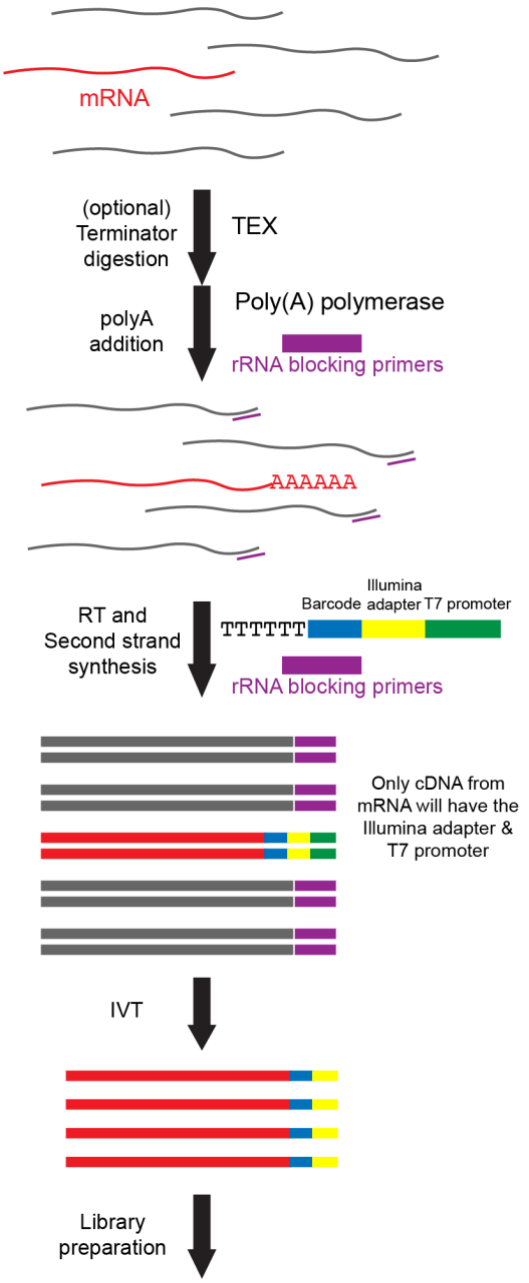


Figure 3.1 Schematic of EMBR-Seq.

After performing an optional Terminator™ 5'-phosphate-dependent exonuclease digestion, poly(A) polymerase and rRNA blocking primers (purple) are added to total bacterial RNA (mRNA in red and rRNA in gray). Blocking primers specifically bind to the 3' end of 5S, 16S, and 23S rRNAs, resulting in the preferential addition of a poly(A)-tail to mRNA molecules. Next, reverse transcription is performed using (i) a poly-T primer, which has an overhang containing a sample-specific barcode (blue), 5' Illumina adapter (yellow), and T7 promoter (green), and (ii) rRNA blocking primers to convert poly-adenylated RNA and rRNA molecules, respectively, to cDNA. The cDNA molecules are then amplified by *in vitro* transcription, and the amplified RNA is used to prepare Illumina libraries. As the rRNA-derived cDNA does not contain a T7 promoter, these molecules are not amplified during *in vitro* transcription, resulting in rRNA depletion.

2. EMBR-seq efficiently depletes rRNA to sequence bacterial mRNA

We applied EMBR-seq to total RNA isolated from the exponential growth phase of *E. coli* strain K12 (MG1655). Starting from 100 ng of total RNA, we were able to successfully make Illumina libraries that were sequenced and mapped to the *E. coli* transcriptome. In parallel, we prepared control libraries where total RNA was processed using the EMBR-seq protocol but in the absence of blocking primers. While total RNA from *E. coli* has previously been reported to consist of 95% rRNA (Giannoukos et al., 2012), our control samples with no blocking primers had approximately 64% rRNA, consistent with previous observations that mRNA molecules are preferentially poly-adenylated compared to rRNA even in the absence of any blocking primers (Figure 3.2A) (Wendisch et al., 2001; Westermann et al., 2016). Importantly, compared to the control samples, we observed a significant increase in rRNA depletion efficiency (from 64% to 16%), with 84% of the mapped reads corresponding to mRNA in samples treated with blocking primers (Figure 3.2A). As tRNAs make up another major class of RNA molecules, we analyzed our data to quantify the detection of these molecules (Westermann et al., 2012). We found tRNA-derived reads to constitute only 0.37% and 1.26% of the mapped reads in the control and EMBR-seq samples, respectively. This expected low detection rate is likely due to the small size of tRNAs, their stable secondary structures and utilization of numerous post-transcriptionally modified nucleotides that are known to interfere with reverse transcription (Cozen et al., 2015; Zheng et al., 2015). Overall, these results demonstrate that EMBR-seq achieves a level of mRNA enrichment that is better or comparable to recent bacterial rRNA depletion reports (Armour et al., 2009; Culviner et al., 2020; He et al., 2010; Huang et al., 2020; Kraus et al., 2019; Petrova et al., 2017; Prezza et al., 2020).

In certain applications, such as those where RNA is extracted from non-cultivable bacterial species within natural isolates, the total RNA can be fragmented and of poor quality. To determine if EMBR-seq can still be successfully applied to degraded RNA, we compared the rRNA depletion efficiency of total *E. coli* RNA with two different RIN (RNA Integrity Number) scores of 7.2 and 2.4. The degraded sample with a RIN score of 2.4 was prepared by heating the total RNA at 95°C for 5 minutes. Surprisingly, we observed that EMBR-seq depleted rRNA to similar levels of 15% and 17% in the untreated and degraded samples, respectively (Appendix 5.D2: Figure S3.1). We hypothesize this occurs because both rRNA and mRNA molecules are fragmented to similar extents, and the increased poly-adenylation of rRNA fragments is matched by a similar increase in poly-adenylation of mRNA fragments, resulting in similar downstream detection of rRNA- and mRNA-derived reads. Thus, these results suggest that EMBR-seq can be effectively applied to sequence the transcriptome of fragmented lower quality total RNA.

We also tested modified blocking primers with a 3' phosphorylation, designed to prevent Superscript II from reverse transcribing rRNA molecules. As expected, we observed rRNA depletion in these samples as well (from 64% to 22%), with 78% of the mapped reads corresponding to mRNA (Figure 3.2A). However, compared to the unmodified blocking primers, these phosphorylated blocking primers were slightly less efficient at rRNA depletion (Figure 3.2A). As the 3' phosphorylated primers prevent polymerase extension, we hypothesize that the reduced rRNA depletion efficiency arises from the small fraction of rRNA molecules that get poly-adenylated, primed by the poly-T primers, and copied through the short 30 bp RNA-DNA hybrid due to the strand-displacement activity of the reverse transcriptase. Therefore, given the reduced efficiency and higher costs of the 3'

phosphorylated blocking primers, all further experiments were performed with unmodified blocking primers.

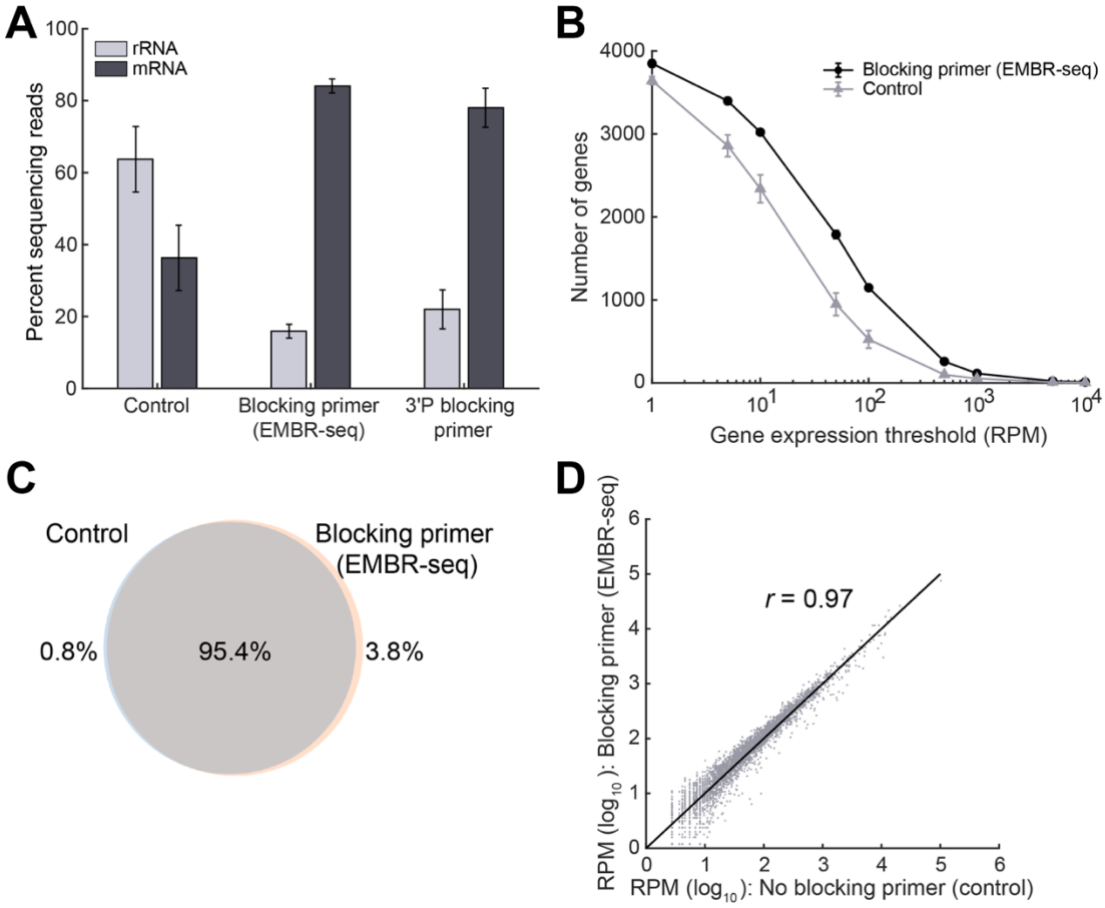


Figure 3.2 Blocking primers in EMBR-seq deplete rRNA and provide a deeper view of the transcriptome without introducing technical biases.

(A) In the presence of blocking primers, a 4-fold rRNA depletion and more than 2-fold mRNA enrichment is achieved compared to control samples. With the introduction of blocking primers in EMBR-seq, mRNAs account for more than 80% of the mapped reads, which is a greater than 16-fold increase compared to total RNA in *E. coli* cells. The 3' phosphorylated blocking primers display similar but slightly lesser mRNA enrichment ($n \geq 2$ replicates for all conditions). (B) Comparison between EMBR-seq and control samples in the number of genes detected above different expression thresholds ($n = 3$ for both conditions). For the EMBR-seq group, error bars are of the same scale as the size of the data points (C) Venn diagram shows that more than 99% of the genes detected in the control samples were also detected when using blocking primers in EMBR-seq. 99.2 % of all detected genes were found in the EMBR-seq samples and 96.2% in the control samples. The number of genes detected were calculated by combining data obtained from three control samples and three EMBR-seq processed samples. (D) Gene transcript counts with and without blocking primers are highly correlated (Pearson $r = 0.97$) suggesting that EMBR-seq does not introduce technical artifacts in quantifying gene expression (n = 3 for both datasets). These experiments were performed starting with 100 ng total RNA from *E. coli.* Error bars in panels (A) and (B) represent standard deviations.

As an alternate strategy, we also incorporated TEX treatment in EMBR-seq as it has previously been shown to specifically degrade rRNAs with 5'-monophosphate ends but not mRNAs that have 5'-triphosphate ends (He et al., 2010; Kang et al., 2011; Kuchina et al., 2019; Sharma et al., 2010). While we again observed rRNA depletion and a corresponding enrichment of mRNA compared to control samples, the effects were less pronounced with a less than 2-fold rRNA depletion, consistent with previous reports (Appendix 5.D2: Figure S3.2) (He et al., 2010; Kuchina et al., 2019). We hypothesize that this reduced efficiency may arise from the additional cleanup step that is necessary prior to treatment with the poly-A polymerase. As a result, we find that blocking primers alone provide the most significant rRNA depletion and mRNA enrichment, and therefore all further experiments were performed without TEX treatment.

3. EMBR-seq provides a detailed view of the transcriptome without introducing technical biases

In designing the steps of EMBR-seq, we wanted to develop a method that is both easily applied and cost-effective. Due to its simplicity, the cost per rRNA depletion reaction in EMBR-seq is ~$0.40, which is at least an order of magnitude lower than other recent rRNA depletion methods and commercial kits (Armour et al., 2009; Culviner et al., 2020; He et al., 2010; Huang et al., 2020; Kraus et al., 2019; Petrova et al., 2017; Prezza et al., 2020) (Appendix 5.D2: Figure S3.3a and Appendix 5.E2: Tables S2.1, S2.2, S2.3). The total cost of EMBR-seq, starting from total bacterial RNA to the final Illumina library, was estimated to be ~$36 per sample. However, the total cost per sample decreases as more samples are multiplexed in the same Illumina library. For example, when 96 samples are multiplexed, the cost per sample drops to ~$20, primarily due to the pooling of samples after second-

strand synthesis that then requires only a single IVT and Illumina library preparation reaction downstream (Appendix 5.D2: Figure S3.3b and Appendix 5.E2: Table S2.2). Thus, EMBR-seq is a simple and cost-effective approach to sequence mRNA from total bacterial RNA.

4. EMBR-seq provides a detailed view of the transcriptome without introducing technical biases

Next, we systematically compared the gene expression profiles obtained from control and rRNA depleted samples to investigate if the use of blocking primers provides a deeper view of the transcriptome without introducing technical artifacts. First, after downsampling sequencing reads to the same depth, we detected 3628 genes in the control samples, while in the mRNA enriched samples we detected 3852 genes, with 99% of the genes in the control samples also detected in the mRNA enriched samples (Figure 3.2B,C). Moreover, at different levels of downsampling, we detected more genes using EMBR-seq compared to the control samples (Appendix 5.D2: Figure S3.4). This suggests that we can measure the genome-wide gene expression landscape in a more cost-effective way using EMBR-seq. Further, the number of genes detected above different expression thresholds was consistently higher for the mRNA enriched samples compared to the control samples (Figure 3.2B). This shows that EMBR-seq is able to detect more genes at different gene expression levels, spanning over three orders of magnitude. Furthermore, we also observed that EMBR-seq derived reads mapped uniformly across the entire length of operons, with modest 3’ and 5’ end bias, suggesting that this method can be used to effectively quantify the expression of genes within operons (Appendix 5.D2: Figure S3.5). Finally, we observed that gene expression between the control and mRNA enriched samples were highly

correlated (Pearson r = 0.97) revealing that the blocking primers do not introduce technical biases in the quantification of gene expression (Figure 3.2D). Collectively, these results demonstrate that our new cost-effective method is able to accurately capture the transcriptome of bacterial cells.

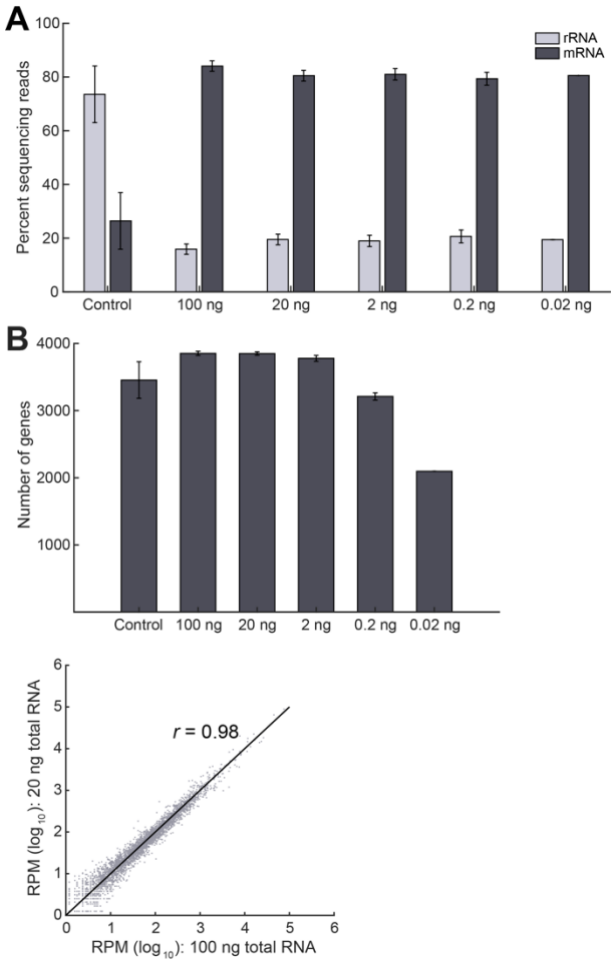


Figure 3.3 EMBR-seq can quantify the transcriptome from low input total RNA.

(A) Similar levels of rRNA depletion and mRNA enrichment are observed when the starting amount of total RNA is decreased from 100 ng to 0.02 ng ($n \geq 2$, except at 0.02 ng where $n = 1$). The control represents average data of control samples made from different input levels of total RNA. The 100 ng data is reproduced from Figure 3.2A. (B) Compared to the control samples, more genes are detected when starting with at least 2 ng input total RNA. Fewer genes are detected when starting total RNA decreases to 0.02 ng. (C) Gene transcript counts are highly correlated (Pearson $r = 0.98$) between 100 ng and 20 ng input total RNA in EMBR-seq. Datasets from lower starting total RNA are also well correlated to the 100 ng samples (Appendix 5.D2: Figure S3.6).

5. EMBR-seq allows mRNA sequencing from low input total RNA

In many practical applications involving non-model and non-cultivable bacterial species, the starting amount of total RNA available for RNA sequencing can be limiting. Therefore, we evaluated if we can successfully deplete rRNA and quantify gene expression from lower amounts of input material. We applied EMBR-seq to 20, 2, 0.2 and 0.02 ng of starting total RNA isolated from the exponential growth phase of *E. coli* strain K12. These starting quantities of total RNA were chosen as they are typically below the sensitivity and detection limit of commercial kits and previously reported methods (Petrova et al., 2017; Prezza et al., 2020). As before, we observed a greater than 3-fold depletion of rRNA across the range of input starting material, including at the lowest starting amount of 0.02 ng total RNA, with greater than 77% of the reads in the sequencing library deriving from mRNA molecules (Figure 3.3A). Similarly, we observed that the total number of genes detected is higher than that in the control samples and is unaffected by the starting input amount of total RNA, except at the lower starting amounts of 0.2 ng and 0.02 ng total RNA (Figure 3.3B). Finally, we also observed that gene expression was highly correlated between different amounts of starting total RNA (Figure 3.3C and Appendix 5.D2: Figure S3.6). These experiments conclusively demonstrate that we can successfully apply EMRB-seq to quantify gene expression from total RNA starting as low as 20 pg.

6. rRNA depletion efficiency of EMBR-seq can be further improved through additional blocking primers

In all the EMBR-seq experiments described above, we observed that 13-22% of the mapped reads still derived from rRNA and therefore, we next attempted to further improve the rRNA depletion efficiency of EMBR-seq. Analyzing the mapped coordinates of the

rRNA-derived reads showed that while the 3’ blocking primers in EMBR-seq effectively depleted rRNA-derived reads compared to control samples from the 3’ end of rRNA molecules, specific “hotspot” regions along the entire length of the 16S and 23S rRNA were disproportionately abundant in the rRNA capture profile (Figure 3.4A,B). We hypothesized that these reads resulted from the combined effects of poly-adenylation of fragmented RNA and biased capture of IVT amplified RNA molecules by random hexamer primers during reverse transcription. This reverse transcription step is part of the final Illumina library preparation protocol where the IVT amplified RNA is first reverse transcribed prior to generation of the Illumina libraries by PCR (Hashimshony et al., 2016; Mooijman et al., 2016; Rooijers et al., 2019). To minimize reads from these specific rRNA regions, we introduced 3 additional blocking primers per rRNA species that targeted the following hotspot locations: coordinates 107, 682, 1241 on 16S rRNA and coordinates 375, 1421, 1641 on 23S rRNA. We found that these hotspot blocking primers successfully reduced rRNA-derived reads from their target locations in the final sequencing library (Figure 4A, B). Overall, this resulted in further improvement in the rRNA depletion efficiency of EMBR-seq with only 10% of the mapped reads deriving from rRNA (Figure 3.4C). These results demonstrate that EMBR-seq is a versatile technique that can be used to effectively deplete rRNA and can potentially be extended to target and deplete any undesired RNA species.

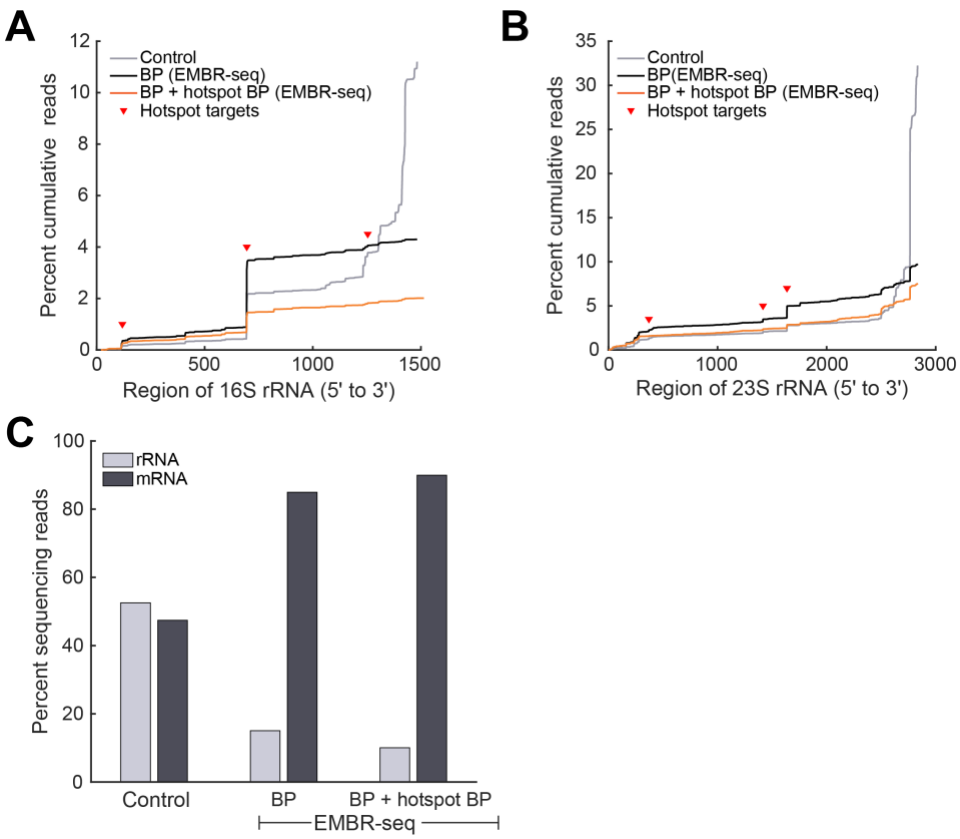


Figure 3.4 Additional hotspot blocking primers increase the rRNA depletion efficiency of EMBR-seq.

(A,B) Cumulative percentage of sequencing reads ordered by mapping location along the (A) 16S and (B) 23S rRNA subunits, from 5' to 3' ends of the transcript. In the control group, the majority of mapped reads are derived from the 3' end together with a few "hotspot" locations (red triangles) along the gene body (gray lines). In EMBR-seq, 3' end blocking primers sharply reduce the number of reads derived from the 3' end (black lines) with the remaining rRNA reads primarily deriving from hotspot locations. Additional blocking primers were designed to minimize poly(A)-tailing and amplification from the vicinity of these coordinates, resulting in further rRNA depletion (orange lines). (C) rRNA depletion and mRNA enrichment is enhanced upon the addition of hotspot blocking primers. With 3' end and hotspot blocking primers in EMBR-seq, mRNA molecules account for 90% of the mapped reads.

C. Discussion

We have developed a new technology, EMBR-seq, to efficiently deplete rRNA from total RNA, thereby enabling a deeper view of the genome-wide distribution of mRNA in bacterial samples. Sequencing bacterial mRNA poses several challenges; for example, the inability to easily enrich mRNA that typically makes up less than 5% of total RNA and the limiting starting amounts of total RNA that may be available when working with non-cultivable bacterial samples (Giannoukos et al., 2012; Mortazavi et al., 2008; Proudfoot,

2011). Through the use of a single 3' blocking primer per rRNA species, EMBR-seq efficiently minimizes the downstream amplification of rRNA molecules, thereby enabling a 4-fold depletion of rRNA in the final sequencing library (Figure 3.1 and 3.2A). In the future, when introducing blocking primers at the 3' end of an rRNA species, we hypothesize that the rRNA depletion efficiency could be further improved by designing primers with a 5' overhang or a bulky 5' modification. As demonstrated in this work, the design of blocking primers at the 3' end of rRNA molecules efficiently depletes rRNA from high quality total RNA samples; however, certain practical applications can produce degraded and fragmented RNA in which rRNA molecules may be less effectively depleted. While we show that the 3' blocking primer alone is sufficient to obtain efficient rRNA depletion when starting with degraded *E. coli* total RNA, a more generalized strategy to overcome this challenge in EMBR-seq is to design additional blocking primers per rRNA species, that span the transcript length to minimize amplification of degraded rRNA molecules (Figure 3.4 and Appendix 5.D2: Figure S3.1).

Starting with total RNA from *E. coli*, we show that efficient depletion of rRNA by EMBR-seq provides higher coverage of the transcriptome at the same sequencing depth (Figure 3.2B and Appendix 5.D2: Figure S3.4). For example, compared to the control samples, the number of unique genes detected increases from 3628 to 3852 in EMBR-seq (Figure 3.2B). In particular, EMBR-seq improves detection of lowly expressed genes below 500 RPM (Figure 3.2B). Further, EMBR-seq provides a more in-depth view of the transcriptional landscape without introducing technical artifacts. We find that 99% of the genes detected in the control group are also detected by EMBR-seq, and that gene expression levels between the two groups are highly correlated (Figure 3.2C, D).

As EMBR-seq typically uses a single blocking primer per rRNA species, sequence conservation analysis of 16S and 23S rRNA suggests that it is adaptable to other microbial species and complex bacterial communities (Appendix 5.D2: Figure S3.7). Recent approaches that employ a large array of probes also achieve a high efficiency of rRNA degradation; however, the need to generate such a large pool of molecules makes it more challenging to extrapolate these methods to evolutionarily distant bacterial species compared to EMBR-seq (Armour et al., 2009; Huang et al., 2020; Prezza et al., 2020). In addition, the use of just one or a few primers per rRNA species combined with the high level of sample multiplexing reduces cost significantly compared to other methods, enabling cost-effective and high-throughput processing of hundreds of samples simultaneously (Appendix 5.D2: Figure S3.3 and Appendix 5.E2: Table S2.1, Table S2.2). Finally, beyond depleting rRNA and enriching for mRNA, the approach used in EMBR-seq can potentially also be used to target other high abundance transcripts in total RNA or used to enrich for non-coding RNA in both prokaryotic and eukaryotic systems (Chao et al., 2012, 2017; Faridani et al., 2016; Sharma et al., 2010).

We also demonstrated that EMBR-seq enables mRNA sequencing of low input RNA samples below the detection limit of commercial kits (Figure 3.3A). Bacterial populations frequently contain diverse species, and even isogenic systems have been shown to display substantial cell-to-cell heterogeneity in gene expression that can give rise to dramatic cellular phenotypes (Avraham et al., 2015; Balázsi et al., 2011; Gefen and Balaban, 2009; Miller-Jensen et al., 2011; Raj and van Oudenaarden, 2008; Russell et al., 2017). Therefore, scaling down bacterial mRNA sequencing techniques to a single-cell level will enable quantification of this variability and provide a better understanding of how transcriptomic

heterogeneity regulates cellular function (Avital et al., 2017; Penaranda and Hung, 2019). Over the last few years, a limited number of approaches have been developed to sequence the transcriptome of single bacterial cells. Early proof-of-concept methods were low throughput techniques that sequenced less than 10 single cells and generally suffered from significant technical noise (Kang et al., 2011; Liu et al., 2019; Wang et al., 2015). More recently, Blattman *et al.* employed combinatorial barcoding to circumvent single cell isolation, enabling high throughput single-cell sequencing of bacterial cell (Blattman et al., 2020). However, this method did not deplete rRNA, resulting in mRNA detection efficiencies of ~2.5-10% (or ~200 mRNA per exponential phase *E. coli* cell). As an alternate approach, Imdahl *et al.* used MATQ-seq to generate sufficient cDNA from individually isolated bacterial cells; however, similar to the previous work, this method also did not deplete rRNA prior to sequencing (Imdahl et al., 2020). In another study, Kuchina *et al.* combined rRNA depletion with combinatorial barcoding to achieve ~5-10% mRNA detection efficiencies in *B. subtilis* (Kuchina et al., 2019). These initial efforts suggest that improved methods could significantly advance single-cell mRNA sequencing in bacteria. EMBR-seq can successfully sequence mRNA from as low as 20 pg of total RNA; therefore, we anticipate that by coupling our rRNA depletion strategy with recent combinatorial barcoding techniques, we will be able to extend EMBR-seq to a single-cell resolution in the future (Blattman et al., 2020; Kuchina et al., 2019).

D. Conclusion

EMBR-seq efficiently depletes rRNA and provides a detailed view of the gene expression landscape within bacterial samples. As EMBR-seq depletes rRNA using a single or a few blocking primer per rRNA species, this new method is easily adaptable to other

microbes as well as an order-of-magnitude cheaper than other reported techniques and commercial kits that frequently use a large array of probes to remove rRNA from total RNA. Finally, EMBR-seq effectively captures the transcriptome from 500-fold lower starting total RNA compared to commercial kits, thereby providing a powerful new approach to investigate gene expression patterns in rare and non-cultivable bacterial species.

4. Improvement and applications of EMBR-seq

With the success of EMBR-seq in *E. coli* cells, where 90% of the sequenced reads mapped to mRNA, we now aim to improve upon the method and apply it to mammalian systems to capture ncRNA.

A. *Improvement of EMBR-seq in E. coli*

Even though EMBR-seq mRNA detection efficiency is comparable to other technologies (depleting rRNA to ~10%), all commercial kits could deplete rRNA to <10%, with a majority reporting <2% rRNA left. Therefore, EMBR-seq can likely be further improved by including an additional orthogonal depletion step. A recent method uses RNase H to deplete rRNA, reporting <5% to ~25% of rRNA left (Huang et al., 2020). Moreover, a new enzyme, *Thermus thermophilus argonaute* (TtAgo), is a programmable DNA-endonuclease which requires a short 5'-phosphorylated single-stranded DNA guide to target its activity to specific corresponding sequence on a DNA substrate. The enzyme introduced one break in the phosphodiester backbone of the complementary substrate sequence (Swarts et al., 2014; Wang et al., 2008). TtAgo depletion concept is similar to that of Cas9 but requires less extensive design. Since Cas9 system has been successfully used to deplete rRNA (Prezza et al., 2020), we hypothesize that TtAgo system can also be employed to deplete rRNA within the EMBR system. As simple design is one of EMBR-seq's advantages, we chose TtAgo over Cas9 and proceeded to explore the addition of RNase H and TtAgo to EMBR-seq.

1. EMBR-RNase H sequencing (EMBR-H-seq)

RNase H is an endoribonuclease that specifically hydrolyzes RNA when hybridized to DNA. To integrate RNase H into EMBR-seq, we added rRNA hybridization primers and

RNase H at the aRNA step. The degraded aRNA is then proceeded as described in EMBR-seq (Figure 4.1) (Wangsanuwat et al., 2020). The rRNA hybridization primers are designed based on the *E. coli* hotspot primers and rationale described in Huang *et al.* (See Appendix 5.C2 for primers) (Huang et al., 2020; Wangsanuwat et al., 2020). There are a few differences between the depletion described in the Huang paper and EMBR-H-seq. Namely, Huang *et al.* depletes rRNA at the total RNA step, while EMBR-H-seq depletes rRNA-derived molecules after IVT at the aRNA step, which means that the length of the RNA being depleted will be shorter in the EMBR-H protocol. Moreover, EMBR-H-seq detects RNA molecules from the 3' end, while Huang protocol detects molecules across the entire length. This could lead to some differences in depletion efficiency depending on the RNA quality and the length distribution of the RNA molecules.

The Huang paper tested two types of RNase H enzymes: a traditional RNase H, which was also used in EMBR-seq, and a thermostable Hybridase, and showed that thermostable RNase H generally outperformed the traditional RNase H (Huang et al., 2020). We performed EMBR-H-seq with both types of RNaseH (Appendix 5.C1).

EMBR-H improves the rRNA depletion of *E. coli* compared to EMBR. On average, EMBR with 3' and hotspot blocking primers could deplete reads derived from rRNA down from 74 % to 6.4 % (Figure 4.2A: E). Using just RNase H led to 7.3 % of final reads coming from rRNA (Figure 4.2A: E-16, E-45). Consistent with Huang's report, we also found that thermostable RNase H outperformed traditional RNase H, although the effect is much less significant (Figure 4.2A: E-16, E-45). With EMBR-H, percentage of rRNA decreased to <3.5 % (Figure 4.2A: ER-16, ER-45). These results show that RNase H could be utilized to

deplete amplified RNA and reach high rRNA depletion level, and combining EMBR with RNase H could further improve depletion results to the level comparable to commercial kits.

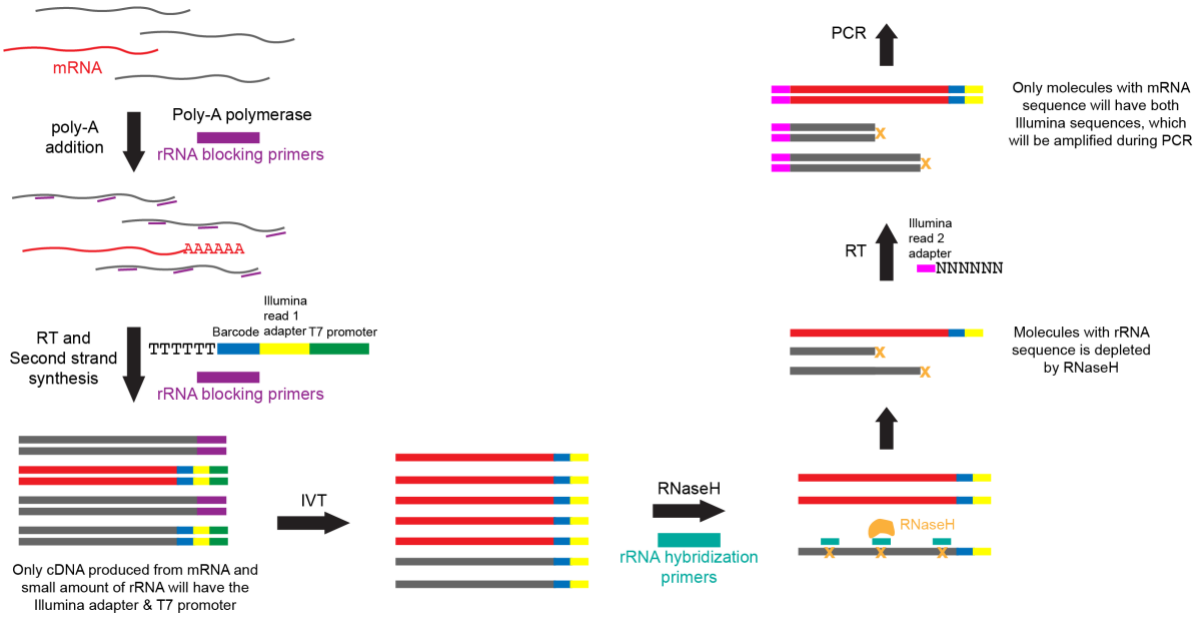


Figure 4.1 Schematic of EMBR-H-seq.

Poly(A) polymerase and rRNA blocking primers (purple) are added to total bacterial RNA (mRNA in red and rRNA in gray). Blocking primers specifically bind to the 3' end of 5S, 16S, and 23S rRNAs as well as their hotspot locations, resulting in the preferential addition of a poly(A)-tail to mRNA molecules. Next, reverse transcription is performed using a poly-T primer and rRNA blocking primers to cDNA. The cDNA molecules are then amplified by *in vitro* transcription. rRNA hybridization primers and RNase H (either traditional or thermostable) are added to the amplified RNA. Then, RNase H selectively degrades molecules derived from rRNA. The remaining RNA is then used to prepare Illumina libraries.

We hypothesize that the inability to significantly deplete further rRNA stems from one or more of the following factors. First, we designed both blocking primers (for EMBR-seq) and rRNA hybridization primers (for RNase H) to target the same rRNA locations, leading to a potential competing depletion. Moreover, as EMBR depleted a significant portion of rRNA-derived molecules by the amplified RNA step, there are likely not many rRNA-derived molecules left for RNase H to deplete later downstream (Figures 3.1, 3.2, 4.1). Lastly, adding an RNase H depletion requires another cleanup step, leading to some material loss, which generally leads to worse depletion result (Figure S4.2).

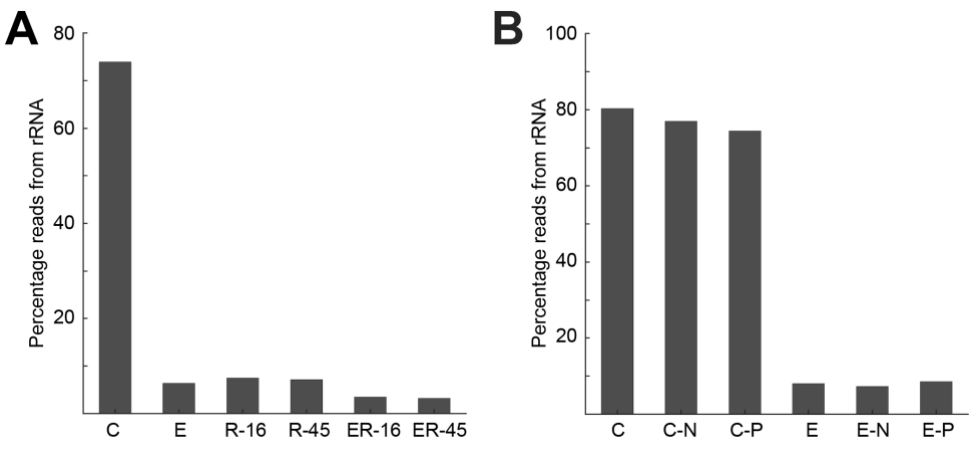


Figure 4.2 EMBR-H and EMBR-T depletion result.

(A,B) The percentage of reads derived from rRNA was plotted for each condition. A) *C* is control, *E* is EMBR-seq, *R-16*, *R-45* are traditional and thermostable RNase H only, respectively, *ER-16* and *ER-45* are EMBR-H with traditional and thermostable RNase H, respectively. B) *C* is control, *E* is EMBR-seq, *-N* and *-P* signify TtAgo treatment without and with pre-PCR respectively ($n = 2$ replicates for all conditions except R-16, R-45, ER-16 and ER-45 where $n = 1$).

2. EMBR-TtAgo sequencing (EMBR-T-seq)

As TtAgo degrades DNA, amplified RNA first needs to be reverse transcribed to cDNA prior to depletion. Moreover, since TtAgo can degrade both single-stranded and double-stranded DNA, two depletion strategies were performed: 1. cDNA product as single-stranded DNA or 2. PCR product as double-stranded DNA (Figure 4.3). TtAgo DNA guide primers for *E. coli* are designed based on manufacturer's recommendation (Appendix 5.C3, 5.C4 for guide primers and full protocol).

Unlike EMBR-H, EMBR-T does not improve rRNA depletion compared to EMBR-seq alone. Using EMBR blocking primers, 91-92% of the mapped reads correspond to mRNA, regardless of whether TtAgo depletion was performed (Figure 4.2B: E, E-N, E-P). Moreover, interestingly, using TtAgo alone does not significantly deplete rRNA, even when compared to control. In control, 20% of reads correspond to mRNA, while 23-25% of reads are derived from mRNA when only TtAgo was used. On the other hand, using RNase H

alone can lead to 93% of reads stemming from mRNA (Figures 4.2). Even though manufacturer's guideline reports that TtAgo performs better on single-stranded substrates (Swarts et al., 2014), we did not notice different depletion levels in the two strategies (Figure 4.2B: C-N, C-P, E-N, E-P).

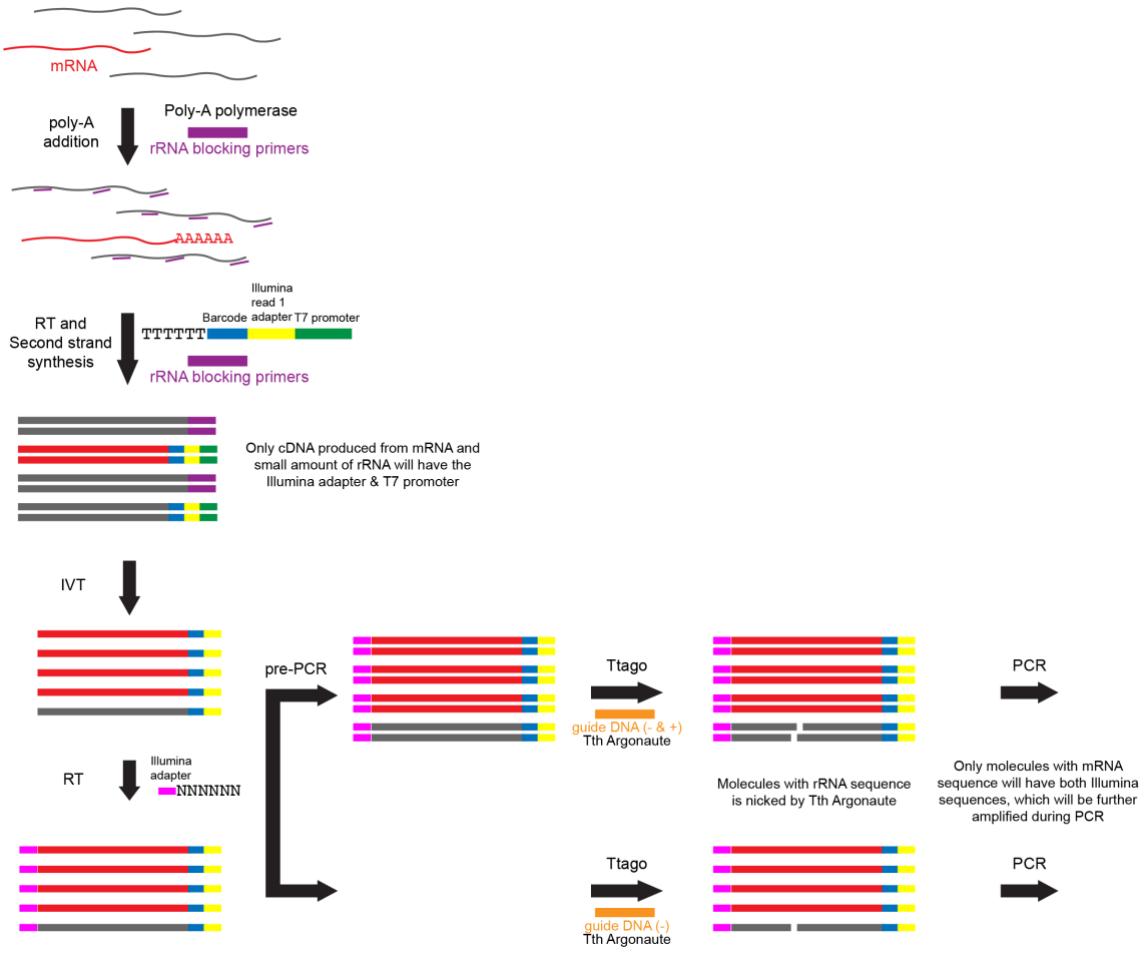


Figure 4.3 Schematic of EMBR-T-seq.

EMBR-T-seq has the same steps as EMBR-seq until the RT step of library preparation. 1) If cDNA target is single-stranded, TtAgo and guide DNA will be added. 2) If cDNA target is double-stranded, cDNA will first undergo 3 cycles of PCR to become double-stranded DNA, followed by TtAgo step, where TtAgo and guide DNA of both strands are added. TtAgo will selectively nick the DNA derived from rRNA, depriving those molecules of having both PCR primers. Only molecules with mRNA sequence will be amplified during PCR.

There are a few potential explanations why TtAgo does not perform well with EMBR. Similar to in the EMBR-H system, TtAgo guide primers are designed for the same hotspot locations; TtAgo was added downstream where there might not be as many molecules to

deplete and requires another cleanup step. In addition, TtAgo enzyme requires temperature to be at least 70°C for maximum activity (Swarts et al., 2014). Incubation at 80°C for 30 minutes can degrade cDNA molecules, including those derived from mRNA, leading to worse depletion results.

Despite preliminary results suggesting that EMBR-seq depletion cannot be significantly improved by an addition of another orthogonal depletion step at the amplified RNA step, we showed that simply adding RNase H along with rRNA hybridization primers, inspired by hotspot blocking primers, at the aRNA step can reach high rRNA depletion level similar to that achieved by EMBR-seq. Moreover, combining EMBR-seq and RNase H to EMBR-H-seq improves the depletion. On the other hand, TtAgo system requires further optimization to achieve comparable rRNA depletion level as other methods reported. Even though both EMBR and RNase H system can reach comparable mRNA enrichment, we generally recommend EMBR over just RNase H because it is simpler and does not require an additional step apart from poly-A addition, when compared to Cel-seq. However, if non-EMBR-depleted amplified RNA has already been made or if further rRNA depletion is desired, RNase H can be added at the aRNA step further downstream. If time and resources allow, we recommend full EMBR-H-seq.

B. *EMBR-seq in mammalian cells*

Current RNA sequencing methods in mammalian cells focus on enriching for mRNA, which possess poly(A)-tail and encode for protein. However, while 90% eukaryotic genomes were transcribed, only about 2% of these transcripts are mRNA (Hubé and Francastel, 2018). The vast majority are transcribed as non-coding RNA (ncRNA).There are many other types of RNAs, e.g. short noncoding RNA or small RNA (sRNA), small

conditional RNA (scRNA), small Cajal body-specific RNA (scaRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), micro RNA (miRNA), miscellaneous RNA (miscRNA), Piwi-interacting RNA (piRNA), small interfering RNA (siRNA), and long non-coding RNA (lncRNA), that are expressed at substantially lower level than mRNA and fulfil regulatory roles and stress response (Dinger et al., 2008; Li et al., 1999; De Santa et al., 2010; West and Greenberg, 2011). Non-coding RNA can be broadly divided into two groups, small ncRNA (shorter than 200 bp) and long ncRNA (longer than 200 bp). A few types of small ncRNA, including the most studied ncRNA, miRNA, have average length of around 20-30 nucleotide. If these very short molecules are captured and have gone through the library preparation steps, the final products will generally have around 140-150 bp in length, which is shorter than the length range normally selected in standard mRNA library (200-1000 bp). While there are many protocols that successfully target small ncRNA, due to their size selection step by gel electrophoresis, these methods naturally lose out on other types of RNA that are outside the target length range (Hadjimichael et al., 2016; Hagemann-Jensen et al., 2018; Jouneau et al., 2012; Wang et al., 2019). As a result, majority of current methods do not capture a full profile of all RNA. The ability to interrogate total RNA content of cells, both short and long, with and without poly(A)-tail, would enable a more complete picture of the transcriptional profile in mammalian cells.

Recently, Smart-seq-total method has been reported to capture total RNA of human primary fibroblasts and induced murine embryonic stem cells differentiated into embryoid bodies (Isakova et al., 2020). Smart-seq-total utilizes *E. coli* poly(A) polymerase to add adenine tails to the 3’ end of RNA molecules and Cas9 protein to remove rRNA (Isakova et al., 2020). EMBR-seq also employs poly(A) polymerase to add adenine tails, but uses the

blocking primers to deplete rRNA at the poly(A) addition step instead of adding a Cas9 depletion step downstream. This leads us to believe that EMBR-seq will also be able to capture total RNA from mammalian cells, without additional rRNA removal step downstream.

1. Strategies for sequencing all RNA in mammalian cells

Because we aim to detect all RNA types present in the cells, some of which have very short length, a standard bead cleanup protocol will remove molecules that are shorter than 200 nucleotides, excluding a lot of short RNA-derived molecules from the samples. Therefore, a right-side bead cleanup, which discards only longer molecules (>1000 bp) and keeps shorter molecules intact, should be performed. However, right-side cleanup will also leave primers and dimers, which have similar length to those short molecules, in the sample solution (~130 bp vs ~150 bp). This could result in poor sequencing results and low mappability, as Illumina sequencer has bias towards shorter reads, leading to a lot of these primers and dimers being sequenced instead of RNA-derived molecules. Therefore, both cleanup strategies: standard cleanup and right-side only cleanup were performed (Appendix 5.C5).

Another challenge for sequencing all RNA is the lack of standard mapping protocol as majority of research focus has been placed on mRNA only. There are two general mapping strategies reported: sequential mapping and annotated genome mapping. In sequential mapping, the sequencing reads were first mapped to one type of RNA, e.g. miRNA, first. The unmapped reads will subsequently be mapped to another type, rRNA/tRNA, then snoRNAs, and so on (Faridani et al., 2016; Jouneau et al., 2012). This strategy will likely inflate the number of reads belonging to the type of RNA that was mapped first, in this case

miRNA, especially when molecules are short and do not map uniquely. The other strategy is to map reads to the entire genome, then GENCODE annotations were used to classify each mapped reads (Hagemann-Jensen et al., 2018; Isakova et al., 2020). One downside of this strategy is that sequencing reads that accidentally map to the unannotated portions of the genome will inflate the mappability percentage. This effect could be significant when the processed reads are short. To be consistent with EMBR-seq data processing pipeline, we created our own reference genome with transcriptome model from RefSeq (O'Leary et al., 2016) and other types of RNA annotated in GENCODE (Frankish et al., 2019) for both mouse and human cells.

2. EMBR-seq in mouse cells

EMBR-seq was performed on 100 ng total RNA of E14 mouse embryonic stem cells with just 3' blocking primers and 3' + hotspot blocking primers (Appendix 5.C5, 5.C6 for primer designs and protocol). As expected, Cel-seq samples essentially do not detect any reads from rRNA (< 0.2%) (Figure 4.4A: C1, C2). On the other hand, adding poly(A)-tail without any blocking primers led to a significant increase in rRNA detection of ~35-40% (Figure 4.4A: N1, N2). Unlike results reported in *E. coli*, adding just 3' blocking primers (EMBR samples: B1-3, B2-3) only slightly depleted rRNA to ~28%. However, adding 3' and hotspot blocking primers (EMBR samples: B1-HS, B2-HS) significantly deplete rRNA-derived reads to ~1%. There is no detectable difference between right-side only cleanup and traditional cleanup, suggesting that the bead cleanup strategy during library preparation is not a significant factor in bulk detection and more samples are likely required to ascertain the effects of having different cleanup strategies.

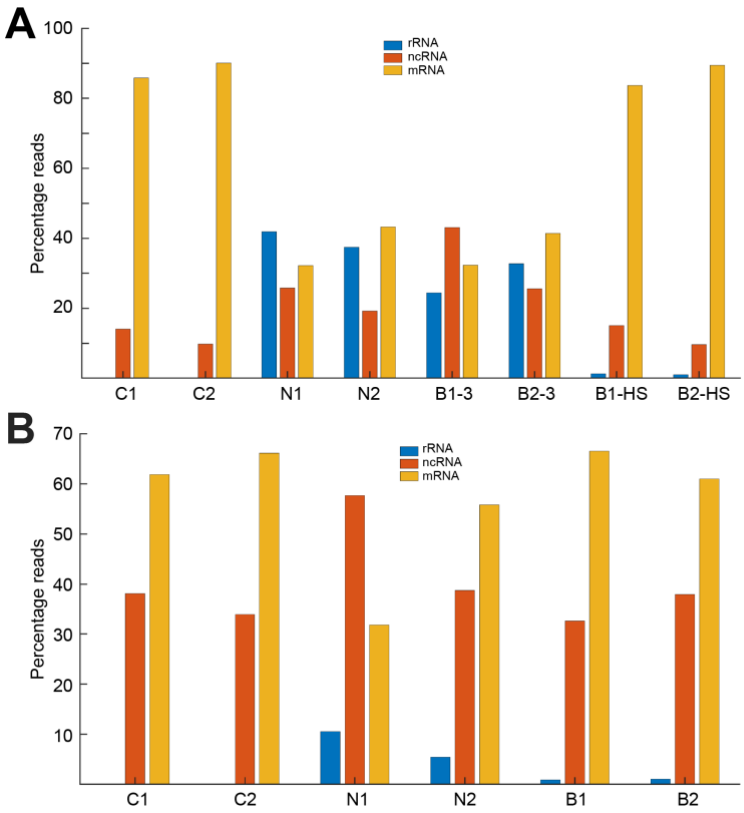


Figure 4.4 EMBR results in mammalian cells.

(A, B) The percentage of reads derived from different types of RNA was plotted for each condition in mouse (A) and human (B) cells. A) *C* is control, *N* is poly-A addition without blocking primers, *B-3* and *B-HS* are EMBR-seq with 3' blocking primers only and with 3' and hotspot blocking primers, respectively. *1* signifies right-side only cleanup, while *2* signifies standard cleanup ($n = 2$ except *B* samples). B) Only EMBR-seq with 3' blocking primers only was performed ($n = 1$ for all samples)

These results suggest that the 3' blocking primers alone might not be able to significantly deplete rRNA and that hotspot primers are also required. Because of poor rRNA depletion in B1-3 and B2-3 cases, the subsequent analysis will only focus on B1-HS and B2-HS cases. Nevertheless, the results show that EMBR protocol can be successfully applied to mammalian cells and the same blocking primer design principle can be extended to other species.

Even though the percentage of reads that map to ncRNA in Cel-seq samples are similar to those in EMBR samples (~12%; Figure 4.4A, ncRNA), we suspect that a significant portion of reads stem from long ncRNA (lncRNA and lincRNA), most of which contain poly(A)-tail (Sun et al., 2018). Therefore, we further classified the reads that mapped to ncRNA to their subtypes and down sampled the mapped reads in each condition to 1 million reads to control for sequencing depth. As expected, EMBR samples have higher percentage of reads mapped to short ncRNA than Cel-seq samples (~0.7% vs ~0.25%; Table 4.1). Moreover, in general, EMBR samples detect more or comparable unique ncRNA genes as Cel-seq samples, without compromising the ability to detect unique mRNA genes (Table 4.1). These results suggest that EMBR-seq protocol can be used to improve short ncRNA detection in mouse cells, leading to a more complete transcriptomic profile.

		Number genes detected					
Condition	% short ncRNA	scaRNA	snRNA	snoRNA	miRNA	miscRNA	mRNA
C1	0.138	4	42.5	59.5	82	28	12379
C2	0.367	4	28	74.5	122	26	12970
B1-HS	0.839	5	65	159	100	33	12452
B2-HS	0.586	3	58	140	115	34	13089

Table 4.1

Percentage of reads derived from short ncRNA and number of different types of RNA for each condition are presented. $n = 2$ for Cel-seq samples and $n = 1$ except for EMBR samples.

3. EMBR-seq in human cells

EMBR-seq was performed on 100 ng total RNA of HEK human embryonic kidney cells with only 3' blocking primers (Appendix 5.C5, 5.C6 for primer designs and protocol). Similar to the results in mouse E14 cells, Cel-seq samples (C1, C2) essentially do not detect any reads from rRNA (<0.02%), while ~5-10% of sequenced reads were derived from rRNA in control samples (N1, N2). Unlike in mouse cells, adding just 3' blocking primers alone (B1, B2) significantly deplete rRNA-derived reads to ~1% (Figure 4.4B). We expect the

depletion percentage to slightly improve if hotspot blocking primers are also added. These results reiterate that EMBR protocol and blocking primer design can be extended to multiple species with similar success.

Similar to the mouse samples, we further classified the reads that mapped to ncRNA to their subtypes and down sampled the mapped reads in each condition to 10 million reads to control for sequencing depth. Surprisingly, EMBR samples have lower percentage of reads mapped to short ncRNA than Cel-seq samples (39.3% vs 41.9%, Table 4.2). However, EMBR samples still detect more or comparable unique ncRNA as in Cel-seq samples, while detecting slightly more unique mRNA genes (Table 4.2). These results suggest that EMBR-seq protocol can also be used to improve short ncRNA detection in human cells, leading to a more complete transcriptomic profile.

Condition	% short ncRNA	scRNA	scaRNA	snRNA	snoRNA	miRNA	miscRNA	mRNA
C1	43.44	1	1	71	39	12	152	15799
C2	40.31	1	0	70	38	16	162	16127
B1	36.93	1	3	141	68	14	270	15841
B2	41.65	1	4	159	70	19	334	16273

Table 4.2

Percentage of reads derived from short ncRNA and number of different types of RNA for each condition are presented. $n = 1$ for all samples.

4. R2 length influences mapping results

In the analysis shown above, we have trimmed Illumina read 2 (R2) to 30 bp and then mapped to the reference file created (Appendix 5.C7). Since some of the small ncRNA are shorter than 30 bp, further trimming R2 down to shorter length could increase the number of reads mapped to short ncRNA.

In B2 case, as R2 was being trimmed to 20 bp, more unique small ncRNA are detected (Table 4.3). There was significant increase in the number of small RNA detected as R2

length decreases from 24 bp to 22 bp, and from 22 bp to 20 bp. However, at 20 bp, B2 no longer detects more unique small ncRNA genes compared to C2, suggesting that higher number of reads that mapped to miRNA might be an artifact of non-unique mapping. Therefore, this result suggests that R2 can be trimmed to 24-25 bp at the shortest, while still maintaining the mapping integrity.

Length	scaRNA	snRNA	snoRNA	miRNA	miscRNA	scRNA	mRNA
20 (C2)	28	1334	418	454	1255	23408	28
20	23	1171	389	360	1204	23253	23
22	16	856	270	203	1043	22667	16
24	11	691	221	120	921	21800	11
25	11	603	197	100	873	21205	11
26	8	569	185	82	845	20754	8
28	7	514	162	74	811	20048	7
30	8	481	151	66	749	19635	8

Table 4.3

Number of unique small ncRNA detected in B2 case with different R2 length. $n = 1$ for all samples.

5. Comparative analysis of EMBR-seq

When the sequencing reads were mapped to just miRNA, the number of unique miRNA detected significantly increase from ~300-400 to ~1700-1800 when R2 was trimmed to 20 bp instead of 30 bp (Table 4.4). Moreover, there are significantly more reads mapped to miRNA when only miRNA was used as a reference, indicating that there is some non-unique mapping (Table 4.4). This result confirms our initial suspicion that sequential mapping greatly inflates the number of the RNA type that was mapped first.

	Full reference		miRNA only	
Length	# genes	% mapped reads	# genes	% mapped reads
C2	454	0.064	1843	3.54
B1	272	0.028	1727	3.46
B2	360	0.074	1843	3.04

Table 4.4

Number of unique miRNA detected in B2 case with different R2 length. $n = 1$ for all samples.

Compared to other methods, EMBR slightly under detected short ncRNA for HEK cells. In Faridani *et al*., they on average captured 3800 miRNA molecules (220 types) and 6500 snoRNA molecules (130 types) in single cells (Faridani et al., 2016). In EMBR standard bead cleanup case (B2) bulk HEK cells down sampled to 10 million reads, 7473 reads are from miRNA (19 types) and 716,529 reads are from snoRNA (70 types). However, as discussed above, sequential mapping likely inflates the number of reads that mapped to miRNA, as miRNA was mapped first. In Smart-seq-total, >100 miRNA types, ~10 scaRNA types, <200 snRNA types, and <100 snoRNA were detected in single cells. On the other hand, EMBR B2 case detected 19 miRNA, 4 scaRNA, 159 snRNA, and 70 snoRNA in bulk. The number of unique short ncRNA will likely decrease when EMBR was applied to single HEK cells. Faridani *et al.* reported that 40% of the miRNA genes detected in 1 μg of RNA can be detected in single cells (Faridani et al., 2016). We expect to detect at least 40% of the short ncRNA genes detected in our bulk 100 ng HEK cells when EMBR is performed in single cells.

In addition, 50% of reads detected in Smart-seq-total are mRNA and 44% are miscRNA (Isakova et al., 2020). Similarly, ~54% reads detected in EMBR samples (B1, B2) are mRNA and EMBR samples also detected significant reads from miscRNA (23%) (Figure 4.5). However, while Smart-seq-total detected ~1% of reads for snRNA and snoRNA each, EMBR samples detected ~9% and ~7% of reads from snRNA and snoRNA (Isakova et al., 2020). These differences in percentage might be due to different starting cells and quantity (bulk vs single cells), different depletion protocols and processing pipeline, or simply technical variability. More in-depth analysis is required to gain insights into the cause of such differences.

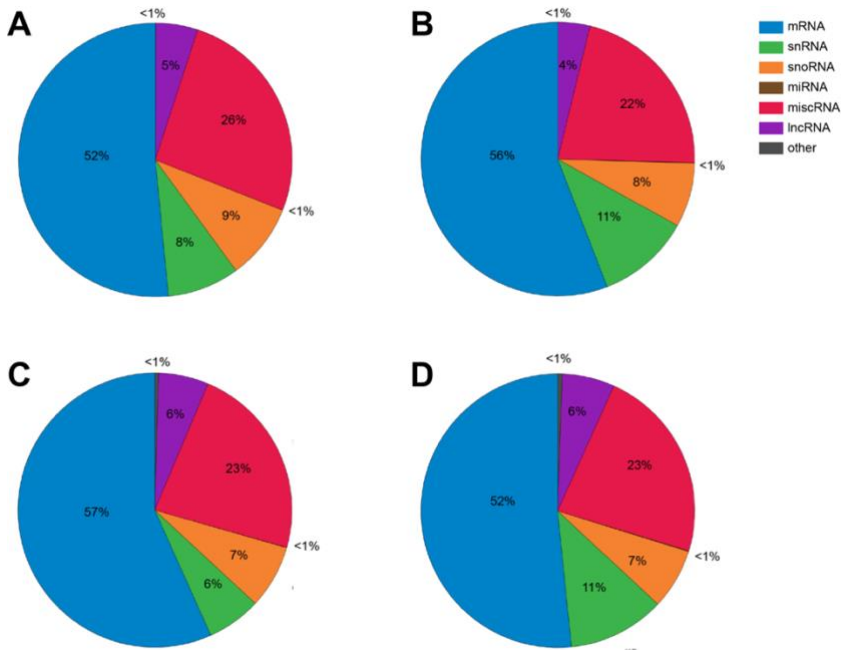


Figure 4.5 Pie chart showing percentage of reads for each RNA type.

For C1 (A), C2 (B), B1 (C), and B2 (D) conditions.

These results suggest that EMBR protocol is able to extract similar information and trends compared to other methods, but further optimization is still required for EMBR to provide comparable detection. Moreover, EMBR protocol should be applied to single HEK cells to show that the protocol is scalable to single cells and to better compare detection results.

5. Concluding remarks

A. *In this dissertation*

The process where an organism develops from a fertilized zygote to a multi-cellular organism with multiple cell types and functions is complex and well-regulated. How the first two distinct cell types emerge in an early embryo is not yet well-understood. In order to study cell differentiation and tissue development, there are two sets of information required: relationships between the cells and cell types. To gain this information, we need the tools to identify cellular lineages and cell types at single cells.

To solve the first part of the problem, in Chapter 2, a short-term endogenous lineage reconstruction technique called scPECLR was presented. With single-cell 5hmC data, scPECLR utilizes how 5hmC is not conserved through cell division and naturally occurring sister chromatid exchange events to infer cellular relationships in early mouse embryos. To expand the reconstruction to larger systems, a single-cell multi-omics method called scH&G-seq that measures 5hmC and genomic DNA was developed. Integrated 5hmC and genomic variants information could significantly improve the prediction accuracy of larger lineage trees. Most importantly, the use of an endogenous epigenetic mark to reconstruct lineage trees at individual cell divisions overcomes major limitations of other lineage reconstruction methods and suggests that scPECLR can be directly extended to study human development.

To solve the second part of the problem, a tool to capture a complete transcriptomic profile of the cells is required. In Chapter 3, a new technology to efficiently deplete rRNA from bacterial total RNA called EMBR-seq was presented. EMBR-seq uses blocking primers to specifically target and deplete rRNA during the poly-adenylation step, resulting in ~90% of sequencing reads derived from mRNA, which originally contributes to <5% in

total RNA. Therefore, EMBR-seq provides higher coverage of the transcriptome at the same sequencing depth. We also demonstrated that EMBR-seq enables mRNA sequencing from as low as 20 pg input total RNA sample, which is below the detection limit of commercial kits, suggesting that this rRNA depletion concept could be used to advance single-cell mRNA sequencing in bacteria.

To further improve the efficiency of rRNA depletion, in Chapter 4, EMBR-seq was combined with an orthogonal depletion method RNase H, resulting in ~96% of sequencing reads derived from mRNA, a level comparable to that using commercial kits. Moreover, due to the simplicity of poly(A) polymerase and blocking primers design, EMBR-seq was extended to mammalian cells to detect ncRNA, which has the same structure as bacterial mRNA. In both mouse and human cells, blocking primers significantly reduce the percentage of reads derived from rRNA compared to control. Moreover, EMBR-seq allows a better detection of unique ncRNA genes at the same sequencing depth.

B. *Outlook and future work*

The dissertation presents methods to reconstruct lineage at high accuracy and to better capture cellular transcriptomic profile. Below lists current challenges and future works required to study tissue development and cell differentiation.

When reconstructing lineages at a single cell-division resolution, the system size will always be one of the limitations. Despite combining 5hmC data with genomic DNA, the system size is likely still limited at 32 cells due to the facts that 5hmC is now distributed among many more cells and there are exponentially more tree topologies to discern (Wangsanuwat et al., 2021). Therefore, to reconstruct lineage in a much larger system such as whole organ or whole organism, scPECLR must be combined with other lineage

reconstruction methods that resolve large-scale clonal information such as those using CRISPR-Cas9. Such methods will initially group cells into different clusters, where scPECLR would further identify cellular relationships within the clusters using single-cell 5hmC. However, as with any large scale Cas9 system, the combined method will unavoidably involve genetic modification and lose endogeneity advantage present in scPECLR. Moreover, in order to resolve lineage information at single cell-division resolution, all single cells in a system must be captured and sequenced, which is very difficult to perform with current technologies.

While EMBR-seq provides a promising rRNA depletion result, both in bacterial and mammalian cells, EMBR-seq must be performed in single cells in order to identify cell types. Due to its simplicity, the concept of poly(A) addition and blocking primers can likely be incorporated into existing single cell RNA sequencing protocols, both in bacterial and mammalian cells (Avital et al., 2017; Hashimshony et al., 2016; Isakova et al., 2020; Penaranda and Hung, 2019).

Lastly, to finally identify cellular lineages and cell types from the same single cells, scAba-Seq/scH&G must be combined with RNA-seq/EMBR-seq to measure both 5hmC and mRNA from the same cells. The 5hmC information will yield the cellular lineages information, while the mRNA, and ncRNA in the case of mammalian cells, can be used to identify cell types. The integrated information will provide insights into how the first two distinct cell types emerge in early embryos and enable us to directly probe symmetric and/or asymmetric cell fate decisions of stem cells at an individual cell division resolution. We anticipated the combined 5hmC/mRNA method will yield significant insights into the tissue

development process, providing applications in both regenerative medicine and stem cell biology.

References

Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J., and Van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. Nature *556*, 108–112.

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M., et al. (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat. Methods *6*, 647–650.

Avital, G., Avraham, R., Fan, A., Hashimshony, T., Hung, D.T., and Yanai, I. (2017). scDual-Seq: Mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing. Genome Biol. *18*, 1–8.

Avraham, R., Haseley, N., Brown, D., Penaranda, C., Jijon, H.B., Trombetta, J.J., Satija, R., Shalek, A.K., Xavier, R.J., Regev, A., et al. (2015). Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses. Cell *162*, 1309–1321.

Balázsi, G., van Oudenaarden, A., and Collins, J.J. (2011). Cellular Decision Making and Biological Noise: From Microbes to Mammals. Cell *144*, 910–925.

Barajas, J.F., Blake-Hedges, J.M., Bailey, C.B., Curran, S., and Keasling, J.D. (2017). Engineered polyketides: Synergy between protein and host level engineering. Synth. Syst. Biotechnol. *2*, 147–166.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422–425.

Biezuner, T., Spiro, A., Raz, O., Amir, S., Milo, L., Adar, R., Chapal-Ilani, N., Berman, V., Fried, Y., Ainbinder, E., et al. (2016). A generic, cost-effective, and scalable cell lineage

analysis platform. Genome Res. *26*, 1588–1599.

Blattman, S.B., Jiang, W., Oikonomou, P., and Tavazoie, S. (2020). Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. Nat. Microbiol.

Bult, C.J., Blake, J.A., Smith, C.L., Kadin, J.A., Richardson, J.E., Anagnostopoulos, A., Asabor, R., Baldarelli, R.M., Beal, J.S., Bello, S.M., et al. (2019). Mouse Genome Database (MGD) 2019. Nucleic Acids Res. *47*, D801–D806.

Chao, Y., Papenfort, K., Reinhardt, R., Sharma, C.M., and Vogel, J. (2012). An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. EMBO J. *31*, 4005–4019.

Chao, Y., Li, L., Girodat, D., Förstner, K.U., Said, N., Corcoran, C., Śmiga, M., Papenfort, K., Reinhardt, R., Wieden, H.-J., et al. (2017). In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways. Mol. Cell *65*, 39–51.

Chatterjee, K., and Wan, Y. RNA.

Claussin, C., Porubský, D., Spierings, D.C.J., Halsema, N., Rentas, S., Guryev, V., Lansdorp, P.M., and Chang, M. (2017). Genome-wide mapping of sister chromatid exchange events in single yeast cells using strand-seq. Elife *6*, 1–17.

Conboy, M.J., Karasov, A.O., and Rando, T.A. (2007). High Incidence of Non-Random Template Strand Segregation and Asymmetric Fate Determination In Dividing Stem Cells and their Progeny. PLoS Biol. *5*, e102.

Cozen, A.E., Quartley, E., Holmes, A.D., Hrabeta-Robinson, E., Phizicky, E.M., and Lowe, T.M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat. Methods *12*, 879–884.

Creecy, J.P., and Conway, T. (2015). Quantitative bacterial transcriptomics with RNA-seq. Curr. Opin. Microbiol. *23*, 133–140.

Culviner, P.H., Guegler, C.K., and Laub, M.T. (2020). A Simple, Cost-Effective, and Robust Method for rRNA Depletion in RNA-Sequencing Studies. MBio *11*.

Dinger, M.E., Amaral, P.P., Mercer, T.R., Pang, K.C., Bruce, S.J., Gardiner, B.B., Askarian-Amiri, M.E., Ru, K., Solda, G., Simons, C., et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. *18*, 1433–1445.

Diroma, M.A., Varvara, A.S., Attimonelli, M., Pesole, G., and Picardi, E. (2020). Investigating Human Mitochondrial Genomes in Single Cells. Genes *11*, 534.

Dvořák, P., Nikel, P.I., Damborský, J., and de Lorenzo, V. (2017). Bioremediation 3 . 0 : Engineering pollutant-removing bacteria in the times of systemic biology. Biotechnol. Adv. *35*, 845–866.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., et al. (2015). Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. Neuron *85*, 49–59.

Falconer, E., Chavez, E.A., Henderson, A., Poon, S.S.S., McKinney, S., Brown, L., Huntsman, D.G., and Lansdorp, P.M. (2010). Identification of sister chromatids by DNA template strand sequences. Nature *463*, 93–97.

Falconer, E., Hills, M., Naumann, U., Poon, S.S.S., Chavez, E. a, Sanders, A.D., Zhao, Y., Hirst, M., and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods *9*, 1107–1112.

Faridani, O.R., Abdullayev, I., Hagemann-Jensen, M., Schell, J.P., Lanner, F., and

Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. Nat. Biotechnol. *34*, 1264.

Feng, Y., and Cohen, S.N. (2000). Unpaired terminal nucleotides and 5' monophosphorylation govern 3' polyadenylation by Escherichia coli poly(A) polymerase I. Proc. Natl. Acad. Sci. *97*, 6415–6420.

Frank, E., and Sanes, J.R. (1991). Lineage of neurons and glia in chick dorsal root ganglia: analysis in vivo with a recombinant retrovirus. Development *111*, 895–908.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773.

Frieda, K.L., Linton, J.M., Hormoz, S., Choi, J., Chow, K.-H.K., Singer, Z.S., Budde, M.W., Elowitz, M.B., and Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. Nature *541*, 107–111.

Gefen, O., and Balaban, N.Q. (2009). The importance of being persistent: heterogeneity of bacterial populations under antibiotic stress. FEMS Microbiol. Rev. *33*, 704–717.

Gell, J.J., Liu, W., Sosa, E., Chialastri, A., Hancock, G., Tao, Y., Wamaitha, S.E., Bower, G., Dey, S.S., and Clark, A.T. (2020). An Extended Culture System that Supports Human Primordial Germ Cell-like Cell Survival and Initiation of DNA Methylation Erasure. Stem Cell Reports *14*, 433–446.

Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Levin, J.Z., Livny, J., Earl, A.M., Gevers, D., Ward, D. V, et al. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. *13*, r23.

Goolam, M., Scialdone, A., Graham, S.J.L., Macaulay, I.C., Jedrusik, A., Hupalowska,

A., Voet, T., Marioni, J.C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell *165*, 61–74.

Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Hadjimichael, C., Nikolaou, C., Papamatheakis, J., and Kretsovali, A. (2016). MicroRNAs for Fine-Tuning of Mouse Embryonic Stem Cell Fate Decision through Regulation of TGF-β Signaling. Stem Cell Reports *6*, 292–301.

Hagemann-Jensen, M., Abdullayev, I., Sandberg, R., and Faridani, O.R. (2018). Small-seq for single-cell small-RNA sequencing. Nat. Protoc. *13*, 2407–2424.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. *17*, 1–7.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R., et al. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat. Methods *7*, 807–812.

Hongslo, J.K., Brøgger, A., Bjørge, C., and Holme, J.A. (1991). Increased frequency of sister-chromatid exchange and chromatid breaks in lymphocytes after treatment of human volunteers with therapeutic doses of paracetamol. Mutat. Res. *261*, 1–8.

Huang, Y., Sheth, R.U., Kaufman, A., and Wang, H.H. (2020). Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. Nucleic Acids Res. *48*, e20–e20.

Hubé, F., and Francastel, C. (2018). Coding and Non-coding RNAs, the Frontier Has Never Been So Blurred. Front. Genet. *9*.

Huelsenbeck, J.P., and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754–755.

Huh, Y.H., Cohen, J., and Sherley, J.L. (2013). Higher 5-hydroxymethylcytosine identifies immortal DNA strand chromosomes in asymmetrically self-renewing distributed stem cells. Proc. Natl. Acad. Sci. *110*, 16862–16867.

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.E., and Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. Nat. Microbiol.

Inoue, A., and Zhang, Y. (2011). Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. Science *334*, 194–194.

Iqbal, K., Jin, S.-G., Pfeifer, G.P., and Szabo, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc. Natl. Acad. Sci.

Isakova, A., Neff, N., and Quake, S.R. (2020). Single cell profiling of total RNA using Smart-seq-total. BioRxiv *21*, 1–9.

Jouneau, A., Ciaudo, C., Sismeiro, O., Brochard, V., Jouneau, L., Vandormael-Pournin, S., Coppée, J.Y., Zhou, Q., Heard, E., Antoniewski, C., et al. (2012). Naive and primed murine pluripotent stem cells have distinct miRNA expression profiles. Rna *18*, 253–264.

Ju, Y.S., Martincorena, I., Gerstung, M., Petljak, M., Alexandrov, L.B., Rahbari, R., Wedge, D.C., Davies, H.R., Ramakrishna, M., Fullam, A., et al. (2017). Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature *543*, 714–718.

Kalhor, R., Mali, P., and Church, G.M. (2017). Rapidly evolving homing CRISPR barcodes. Nat. Methods *14*, 195–200.

Kang, Y., Norris, M.H., Zarzycki-Siek, J., Nierman, W.C., Donachie, S.P., and Hoang, T.T. (2011). Transcript amplification from single bacterium for transcriptome analysis. Genome Res. *21*, 925–935.

Karpowicz, P., Morshead, C., Kam, A., Jervis, E., Ramuns, J., Cheng, V., and Van Der Kooy, D. (2005). Support for the immortal strand hypothesis: Neural stem cells partition DNA asymmetrically in vitro. J. Cell Biol. *170*, 721–732.

Kraus, A.J., Brink, B.G., and Siegel, T.N. (2019). Efficient and specific oligo-based depletion of rRNA. Sci. Rep. *9*, 1–8.

Kretzschmar, K., and Watt, F.M. (2012). Lineage tracing. Cell *148*, 33–45.

Kuchina, A.A., Brettner, L.M., Paleologu, L., and Roco, C.M. (2019). Microbial single-cell RNA sequencing by split-pool barcoding. 1–20.

Kung, Y., Runguphan, W., and Keasling, J.D. (2012). From Fields to Fuels: Recent Advances in the Microbial Production of Biofuels. ACS Synth. Biol. *1*, 498–513.

Li, T.-H., Spearow, J., Rubin, C.M., and Schmid, C.W. (1999). Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. Gene *239*, 367–372.

Liu, Q., Jiang, C., Xu, J., Zhao, M.-T., Van Bortle, K., Cheng, X., Wang, G., Chang, H.Y., Wu, J.C., and Snyder, M.P. (2017). Genome-Wide Temporal Profiling of Transcriptome and Open Chromatin of Early Cardiomyocyte Differentiation Derived From hiPSCs and hESCs. Circ. Res. *121*, 376–391.

Liu, Y., Jeraldo, P., Jang, J.S., Eckloff, B., Jen, J., and Walther-Antonio, M. (2019). Bacterial Single Cell Whole Transcriptome Amplification in Microfluidic Platform Shows Putative Gene Expression Heterogeneity. Anal. Chem.

Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56–62.

Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science *350*, 94–98.

Lodish, H., Berk, A., and Zipursky, S. (2000a). Section 11.6, Processing of rRNA and tRNA. In Molecular Cell Biology., (New York: W. H. Freeman), p.

Lodish, H., Berk, A., and Zipursky, S. (2000b). Section 11.2, Processing of Eukaryotic mRNA. In Molecular Cell Biology, p.

LoTurco, J., Manent, J.-B., and Sidiqi, F. (2009). New and Improved Tools for In Utero Electroporation Studies of Developing Cerebral Cortex. Cereb. Cortex *19*, i120–i125.

Ludwig, L.S., Lareau, C.A., Ulirsch, J.C., Christian, E., Muus, C., Li, L.H., Pelka, K., Ge, W., Oren, Y., Brack, A., et al. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. Cell *176*, 1325–1339.

Madeira, F., Park, Y. mi, Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. *47*, W636–W641.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

Messerschmidt, D.M., Knowles, B.B., and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplanatation embryos. Genes Dev. *28*, 812–828.

Miller-Jensen, K., Dey, S.S., Schaffer, D. V., and Arkin, A.P. (2011). Varying virulence: epigenetic control of expression noise and disease processes. Trends Biotechnol. *29*, 517–525.

Mooijman, D., Dey, S.S., Boisset, J.-C., Crosetto, N., and van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. Nat. Biotechnol. *34*, 852–856.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature *496*, 229–232.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

Otero, J.M., and Nielsen, J. (2010). Industrial systems biology. Biotechnol. Bioeng. *105*, 439–460.

Pei, W., Feyerabend, T.B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., et al. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature *548*, 456–460.

Penaranda, C., and Hung, D.T. (2019). Single-Cell RNA Sequencing to Understand Host–Pathogen Interactions. ACS Infect. Dis. *5*, 336–344.

Peng, X. "Nick," Gilmore, S.P., and O'Malley, M.A. (2016). Microbial communities for bioprocessing: lessons learned from nature. Curr. Opin. Chem. Eng. *14*, 103–109.

Perli, S.D., Cui, C.H., and Lu, T.K. (2016). Continuous genetic recording with self-targeting CRISPR-Cas in human cells. Science *353*.

Petrova, O.E., Garcia-alcalde, F., Zampaloni, C., and Sauer, K. (2017). Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. Nat. Publ. Gr. 1–15.

Potten, C.S., Owen, G., and Booth, D. (2002). Intestinal stem cells protect their genome by selective segregation of template DNA strands. J. Cell Sci. *115*, 2381–2388.

Prezza, G., Heckel, T., Dietrich, S., Homberger, C., Westermann, A.J., and Vogel, J. (2020). Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. RNA *26*, 1069–1078.

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes Dev. *25*, 1770–1782.

Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.

Rocheteau, P., Gayraud-Morel, B., Siegl-Cachedenier, I., Blasco, M.A., and Tajbakhsh, S. (2012). A subpopulation of adult skeletal muscle stem cells retains all template DNA

strands after cell division. Cell *148*, 112–125.

Rooijers, K., Markodimitraki, C.M., Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey, S.S., and Kind, J. (2019). Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. Nat. Biotechnol. *37*, 766–772.

Russell, J.R., Cabeen, M.T., Wiggins, P.A., Paulsson, J., and Losick, R. (2017). Noise in a phosphorelay drives stochastic entry into sporulation in Bacillus subtilis. EMBO J. *36*, 2856–2869.

Saenko, S. V, French, V., Brakefield, P.M., and Beldade, P. (2008). Conserved developmental processes and the formation of evolutionary novelties: examples from butterfly wings. Philos. Trans. R. Soc. B Biol. Sci. *363*, 1549–1556.

Saitou, M., Kagiwada, S., and Kurimoto, K. (2012). Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. Development *139*, 15–31.

Salipante, S.J., Kas, A., McMonagle, E., and Horwitz, M.S. (2010). Phylogenetic analysis of developmental and postnatal mouse cell lineages. Evol. Dev. *12*, 84–94.

Sanders, A.D., Meiers, S., Ghareghani, M., Porubsky, D., Jeong, H., van Vliet, M.A.C.C., Rausch, T., Richter-Pechańska, P., Kunz, J.B., Jenni, S., et al. (2020). Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. Nat. Biotechnol. *38*, 343–354.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. PLoS Biol. *8*, e1000384.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J.S., Alquicira-Hernández, K., Muñiz-Rascado,

L.J., Peña-Loredo, P., et al. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. *47*, D212–D220.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. Nature *464*, 250–255.

Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D., et al. (2010). Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. Cell *143*, 134–144.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. Nat. Biotechnol. *36*, 469–473.

Stoddard, S.F., Smith, B.J., Hein, R., Roller, B.R.K., and Schmidt, T.M. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. *43*, D593–D598.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O., and Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. Nature *514*, 322–327.

Sun, Q., Hao, Q., and Prasanth, K. V. (2018). Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet. *34*, 142–157.

Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., Brouns, S.J.J., et al. (2014). DNA-guided DNA interference by a

prokaryotic Argonaute. Nature *507*, 258–261.

Tateishi, S., Niwa, H., Miyazaki, J.-I., Fujimoto, S., Inoue, H., and Yamaizumi, M. (2003). Enhanced Genomic Instability and Defective Postreplication Repair in RAD18 Knockout Mouse Embryonic Stem Cells. Mol. Cell. Biol. *23*, 474–481.

Turner, D.L., and Cepko, C.L. (1987). A common progenitor for neurons and glia persists in rat retina late in development. Nature.

Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science *360*, 981–987.

Wang, J., Chen, L., Chen, Z., and Zhang, W. (2015). RNA-seq based transcriptomic analysis of single bacterial cells. Integr. Biol. *7*, 1466–1476.

Wang, N., Zheng, J., Chen, Z., Liu, Y., Dura, B., Kwak, M., Xavier-Ferrucio, J., Lu, Y.C., Zhang, M., Roden, C., et al. (2019). Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation. Nat. Commun. *10*, 1–12.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D.J. (2008). Structure of the guide-strand-containing argonaute silencing complex. Nature *456*, 209–213.

Wangsanuwat, C., Heom, K.A., Liu, E., O'Malley, M.A., and Dey, S.S. (2020). Efficient and cost-effective bacterial mRNA sequencing from low input samples through ribosomal RNA depletion. BMC Genomics *21*, 717.

Wangsanuwat, C., Chialastri, A., Aldeguer, J.F., Rivron, N.C., and Dey, S.S. (2021). A probabilistic framework for cellular lineage reconstruction using integrated single-cell 5-hydroxymethylcytosine and genomic DNA sequencing. Cell Reports Methods 100060.

Wen, L., and Tang, F. (2018). Boosting the power of single-cell analysis. Nat. Biotechnol. *36*, 408–409.

Wendisch, V.F., Zimmer, D.P., Khodursky, A., Peter, B., Cozzarelli, N., and Kustu, S. (2001). Isolation of Escherichia coli mRNA and Comparison of Expression Using mRNA and Total RNA. *213*, 205–213.

West, A.E., and Greenberg, M.E. (2011). Neuronal Activity-Regulated Gene Transcription in Synapse Development and Cognitive Function. Cold Spring Harb. Perspect. Biol. *3*, a005744–a005744.

Westermann, A.J., Gorski, S.A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. Nat. Rev. Microbiol. *10*, 618–630.

Westermann, A.J., Förstner, K.U., Amman, F., Barquist, L., Chao, Y., Schulte, L.N., Müller, L., Reinhardt, R., Stadler, P.F., and Vogel, J. (2016). Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. Nature *529*, 496–501.

Wetterstrand, K.A. The Cost of Sequencing a Human Genome.

White, M.D., Angiolini, J.F., Alvarez, Y.D., Kaur, G., Zhao, Z.W., Mocskos, E., Bruno, L., Bissiere, S., Levi, V., and Plachta, N. (2016). Long-Lived Binding of Sox2 to DNA Predicts Cell Fate in the Four-Cell Mouse Embryo. Cell *165*, 75–87.

Woodworth, M.B., Girskis, K.M., and Walsh, C.A. (2017). Building a lineage from single cells: genetic techniques for cell lineage tracking. Nat. Rev. Genet. *18*, 230–244.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat. Commun. *2*, 241.

Wu, X., Inoue, A., Suzuki, T., and Zhang, Y. (2017). Simultaneous mapping of active

DNA demethylation and sister chromatid exchange in single cells. Genes Dev. 1–13.

Xu, J., Nuno, K., Litzenburger, U.M., Qi, Y., Corces, M.R., Majeti, R., and Chang, H.Y. (2019). Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. Elife *8*, 1–14.

Zack, G.W., Rogers, W.E., and Latt, S.A. (1977). Automatic measurement of sister chromatid exchange frequency. J. Histochem. Cytochem. *25*, 741–753.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods *12*, 835–837.

Appendix

A. *Chapter 2 Methods*

1. Embryo isolation and cell picking

Embryos were gently flushed out of the infundibulum of E2.5 pregnant mice using warm M2 medium. Embryos were then manipulated in 4-ring IVF dishes coated with RNase-free BSA. Embryos were washed in PBS-0 and in Tyrode's acid to remove the zona pellucida, then placed in a 1/3 dilution of TrypLE Select Gibco A12177-01 (stock solution is referred by Gibco as 10x concentrated) and placed on the warm plate for 2 minutes. Glass capillaries of different diameters were then used to dissociate the embryo into 2-3 clusters. Cells were then progressively extracted from each cluster, one after the other, using glass capillaries. Every single cell that is released from the clusters is immediately placed into a well of a 384-well plate containing lysis buffer.

2. Cell culture and cell sorting

H9 cells were grown on Matrigel (Fisher cat #08-774-552) in mTeSR1 (Stem Cell Technologies cat #85850). Cells were passed in clumps using Versene solution (Thermo fisher scientific cat # 15040066). For sorting, cells were dissociated into single cells using TrypLE, resuspended in 1x PBS, and passed through a cell strainer.

3. Single-cell 5hmC sequencing (scAba-Seq)

Single cells isolated from 4-, 8- and 16-cell mouse embryos were deposited into 384-well plates and the scAba-Seq protocol was performed using the Nanodrop II liquid-handing robot. Briefly, after protease treatment to strip off chromatin, 5hmC sites in the genome

were glucosylated using T4-Phage β-glucosyltransferase. Next, AbaSI, which recognizes glucosylated sites and introduces double-stranded breaks with 3' overhangs 11-13 nucleotides downstream of the recognition site, was added to the reaction mixture. The fragmented genomic DNA molecules were ligated to double-stranded adapters containing a cell barcode, 5' Illumina adapter, and T7 promoter. The ligated molecules were amplified by *in vitro* transcription and then used to prepare Illumina libraries. A detailed protocol can be found in Mooijman et al., 2016.

4. Modeling SCE events as a Poisson process

The 5hmC data was discretized into 2 or 4 Mb bins and all SCE transitions in the 8-cell mouse embryos were identified manually. A specific SCE transition on chromosome 14 was found at the same genomic position in all embryos due to a misorientation of the reference genome (mm10), consistent with previous reports (Falconer et al., 2012; Wu et al., 2017). The stochastic nature of SCE events is modeled as a Poisson process. In using a Poisson process to model SCE events, we assume that all SCE events occur independently and at a constant rate. The probability of observing x SCE transitions in one cell cycle is given by:

$$P[x] = \frac{b^x * e^{-b}}{x!} \quad (1)$$

where b is the average number of SCE transitions per chromosome per cell division. Further, to build a probabilistic framework to reconstruct cellular lineages, we define the following parameters: (1) r is the probability that an original strand is inherited by a particular daughter cell, which is equal to ½ for randomly segregating DNA strands; (2) k_{ij} is the genomic length fraction of the j^{th} segment ($1 \leq j \leq l+1$, where l is the number of

SCE transitions) of the original DNA strand that is observed in cell i; and (3) N is the number of unique positions where SCE events can occur.

5. scPECLR

The first step is to use the numbers of observed SCE events to estimate b using maximum likelihood estimation (MLE). Thereafter, Original Strand Segregation (OSS) analysis is used to separate the cells into two groups, reducing the number of cell divisions to be reconstructed from n to $n-1$. Next, within each subtree, we calculate the probability of observing a SCE pattern of a chromosome given a tree topology. For example, for the most frequently occurring pattern of one SCE event shared between two cells (see example in Figure 2.2A), the probability of observing it in Tree A is given by the product of the probability of having no SCE events in the first cell division and the probability of having one SCE event in the second cell division

$$P(k_{11}, k_{22}|\tau_A) = P_{\tau_A} = \left(\frac{r}{e^b}\right)\left(\frac{br}{e^b N}\right) \tag{2}$$

Similarly, the probability of observing this pattern in Tree B is given by the product of the probability of having one SCE event in the first cell division, and no SCE events within the original DNA strands in both cells in the second cell division

$$\begin{aligned} &P(k_{11}, k_{22}|\tau_B) = P_{\tau_B} \\ &= \left(\frac{br}{e^b N}\right)\left(\frac{r}{e^b} e^{\left(\frac{(1-k_{11})(N+1)}{N}\right)b}\right)\left(\frac{r}{e^b} e^{\left(\frac{(1-k_{22})(N+1)}{N}\right)b}\right) \end{aligned} \tag{3}$$

which leads to

$$\frac{P_{\tau_A}}{P_{\tau_B}} = \frac{2}{e^{\frac{b}{N}}} \tag{4}$$

Detailed analytical expressions for the probability of observing different SCE patters are provided in the Quantification and Statistical Analysis section "Analytical expressions for the probability of observing the three most common SCE patterns".

Subsequently, we assume that the SCE patterns on each chromosome are independent and compute the overall probability of observing SCE events over the whole genome for each tree topology. Moreover, as a 4-cell subtree has only three distinct topologies, we get

$$P(\tau_A|D) + P(\tau_B|D) + P(\tau_C|D) = 1 \quad (5)$$

where D represents the genome-wide SCE patterns in all cells of the embryo. Rearrangement gives us the probability of observing different tree topologies given the SCE patterns over the whole genome

$$P(\tau_A|D) = \frac{1}{1 + \frac{P(D|\tau_B)}{P(D|\tau_A)} + \frac{P(D|\tau_C)}{P(D|\tau_A)}} \quad (6)$$

Finally, the probability of observing the topology of a particular 8-cell tree is a product of the probabilities of the two corresponding 4-cell subtrees.

In 8- and 16-cell predictions, after the probabilities of all tree topologies are estimated, scPECLR assigns the topology with the highest probability as the predicted tree. Then, starting with this predicted tree, b values specific to each cell division are estimated. A second iteration with cell division-specific b values is then performed to obtain a new predicted tree. If the new predicted tree is not the same tree as that inferred in the first iteration, another iteration is performed starting from the predicted tree in the current iteration. This iterative process is carried out till the predicted tree is the same as that obtained in the previous iteration or until 10 iterations have been performed. In all *in vivo* mouse embryos and almost all simulated embryos, the predicted tree converges by the 3rd

iteration. Since we know that the iterative prediction is mostly useful when the rates of SCE events generating the simulated embryos are different for each cell division (see “scPECLR is robust to initial estimates of the SCE rate and to varying SCE rates at each cell division”), iterative prediction was not performed in 32-cell tree predictions to conserve computational resources.

6. scH&G-seq

384-well plates containing 4 μL of Vapor-Lock (Qiagen) and 200 nL of lysis buffer (0.0875% IGEPAL CA-630) are prepared and single cells are FACS sorted into each reaction well. After sorting, plates are stored at -80°C until use. The cells were lysed at 65°C for 3 minutes, and reaction wells receive 500 nL of either 1.4x Buffer4 (NEB) [negative control, scAba-Seq only], BseRI mix [1.4x Buffer 4, 0.25 units BseRI (NEB)], AluI mix [1.4x Buffer 4, 0.25 units AluI (NEB)], or a combined mixture containing both BseRI and AluI (1.4x Buffer 4, 0.125 units BseRI, 0.125 units AluI). BseRI was selected because it yields the same 2 nucleotide 3’ overhang as AbaSI, while AluI was selected because we have previously used it successfully to digest gDNA (Rooijers et al., 2019). The plate is incubated for 1 hour at 37°C followed by heat inactivation at 80°C for 20 minutes. Next, 1.8 μL of protease mix (1x Buffer 4, 6 μg Qiagen protease) is added, and the plate is heated to 50°C for 16 hours, 75°C for 20 minutes, and 80°C for 5 minutes. Then 5hmC sites in the genome are glucosylated by adding 500 nL of glucosylation mix [1x Buffer 4, 1x UDP-Glucose (NEB), 1 unit T4-BGT (NEB)] and incubated at 37°C for 16 hours. Afterwards, 500 nL of protease mix (1x Buffer 4, 2 μg Qiagen protease) is added, and the plate is heated to 50°C for 3 hours, 75°C for 20 minutes, and 80°C for 5 minutes. To detect 5hmC, 500 nL of AbaSI reaction mix (1x Buffer 4, 1 unit AbaSI) is added and the plate is incubated at 25°C

for 1.5 hours, and 65°C for 25 minutes. Cells receiving the AluI mix or the combined BseRI and AluI mix have 200 μL of 64 nM blunt end adapter added as described previously (Rooijers et al., 2019). All cell also receive 200 μL of 75 nM scAba-seq adapters as described in Mooijman et al., 2016. Ligation mix [1x T4 DNA Ligase reaction buffer (NEB), 4 mM ATP (NEB), 140 units T4 DNA Ligase (NEB)] is then added to bring the total volume of each reaction well to 5 μL. Subsequently, the plate is incubated at 16°C for 16 hours. Excluding the Vapor-Lock, all reagents are dispensed using the Nanodrop II liquid handling robot. After ligation, the reaction wells are pooled and the downstream steps are performed as described previously (Gell et al., 2020).

Reads were separated by their molecule type barcodes and mapped to hg19 using Burrows-Wheeler Aligner (BWA). AluI based reads were identified as described in Rooijers et al., 2019. 5hmC based reads were identified as described in Mooijman et al., 2016, with the following modification. A custom Perl script was written to identify if a read also contained a BseRI recognition site. If a read contained recognition sites for both BseRI and AbaSI, it was discarded.

7. Analytical expressions for the probability of observing the three most common SCE patterns

Case I: The most common SCE pattern that we observed in mouse embryos is one SCE transition shared between two cells (cells 1 and 2 in Figure 2.2A and Appendix D1: Figure S2.1). This pattern alone cannot discriminate between sister (Tree A) or cousin (Trees B and C) cell configurations as all three topologies are consistent with the SCE pattern. Therefore, we developed a model to rigorously determine the probability of observing any SCE pattern given a tree topology. For Tree A, the probability of observing one shared SCE transition is

given by the product of the probability of having no SCE events in the first cell division and the probability of having one SCE event in the second cell division. Further, there is a $^1/_N$ chance that the observed SCE event occurs at a specific discretized genomic position. The probability that the original DNA strand is inherited by the mother of cells 1 and 2 is r, and the probability of inheriting the observed SCE pattern between cells 1 and 2 is given by r.

$$P(k_{11}, k_{22}|\tau_A) = P_{\tau_A} = \left(\frac{r}{e^b}\right)\left(\frac{br}{e^b N}\right) \tag{7}$$

Similarly, for Tree B,

$$P(k_{11}, k_{22}|\tau_B) = P_{\tau_B} = \left(\frac{br}{e^b N}\right)\left(\frac{r}{e^b} + m\right)\left(\frac{r}{e^b} + m\right) \tag{8}$$

Here, m represents the probability that the SCE events during the second cell division occur within newly synthesized DNA strands that contain undetectable levels of 5hmC. To estimate m on the left branch of the lineage tree that gives rise to cells 1 and 3, we can show that

Probability of 1 undetectable SCE transition $= \frac{br}{e^b}(1 - k_{11})(\frac{N+1}{N})$

Probability of 2 undetectable SCE transitions $= \frac{b^2 r}{e^b 2!}\left[(1 - k_{11})\left(\frac{N+1}{N}\right)\right]^2$

Probability of n undetectable SCE transitions $= \frac{b^n r}{e^b n!}\left[(1 - k_{11})\left(\frac{N+1}{N}\right)\right]^n$

Therefore, m is given by

$$\begin{aligned} \mathrm{m} &= \frac{br}{e^b}K_N + \frac{b^2 r}{e^b 2!}(K_N)^2 + \frac{b^3 r}{e^b 3!}(K_N)^3 + \cdots \\ &= \frac{r}{e^b}\left(\frac{(K_N b)^1}{1!} + \frac{(K_N b)^2}{2!} + \frac{(K_N b)^3}{3!} + \cdots\right) \\ &= \frac{r}{e^b}(e^{K_N b} - 1) \end{aligned} \tag{9}$$

where $K_N = (1 - k_{11})(\frac{N+1}{N})$.

Thus, (8) becomes

$$P(k_{11}, k_{22}|\tau_B) = P_{\tau_B}$$
$$= \left(\frac{br}{e^b N}\right)\left(\frac{r}{e^b} e^{\left(\frac{(1-k_{11})(N+1)}{N}\right)b}\right)\left(\frac{r}{e^b} e^{\left(\frac{(1-k_{22})(N+1)}{N}\right)b}\right) \quad (10)$$

Further, it is trivial to show that the probability of observing the SCE pattern given Tree B or C is equal, that is

$$P_{\tau_B} = P(k_{11}, k_{22}|\tau_B) = P(k_{11}, k_{22}|\tau_C) = P_{\tau_C} \quad (11)$$

Therefore, the ratio of the probability of cells 1 and 2 being sisters (Tree A) *vs.* cousins (Trees B or C) is given by

$$\frac{P_{\tau_A}}{P_{\tau_B}} = \frac{P_{\tau_A}}{P_{\tau_C}} = \frac{P_{\tau_{sisters}}}{P_{\tau_{cousins}}} = \frac{2}{e^{\frac{b}{N}}} \quad (12)$$

Note that the probability ratio is a function of only the SCE rate and the number of bins, and is not dependent on the location of the SCE event in this case.

Case II: Another common SCE pattern is the observation of two SCE transitions that are shared between two cells (Figure 2.2C and Appendix D1: Figures S2.1, S2.2A). For the original DNA strand to be observed in only two cells, SCE transitions must occur in the same cell cycle. Thus, the probability of observing this SCE pattern in Tree A is given by

$$P(k_{11}, k_{22}, k_{13}|\tau_A) = P_{\tau_A} = (\frac{r}{e^b})(\frac{b^2 r}{e^b 2!}\frac{2}{N^2}) \quad (13)$$

The first term is the probability that no SCE event occurs in the first cell division, and the second term is the probability of having two SCE transitions during the second cell division.

Similarly, for Tree B

$$P(k_{11}, k_{22}, k_{13}|\tau_B) = P_{\tau_B}$$

$$= \left(\frac{b^2 r}{e^b 2!}\frac{2}{N^2}\right)\left(\frac{r}{e^b} + \mathrm{q}\right)\left(\frac{r}{e^b} e^{\left(\frac{(k_{11}+k_{13})(N+1)}{N}\right)b}\right) \tag{14}$$

where q is the probability that undetectable SCE events occur within the 5hmC-depleted genomic region between k_{11} and k_{13}, whose length is equal to k_{22}. Note that the observed SCE pattern is possible for an even number of SCE events occurring within this region. To estimate q, we can show that

Probability of 2 undetectable SCE transitions $= \frac{b^2 r}{e^b 2!}\left[\frac{k_{22}(N+1)+1}{N}\right]^2$

Probability of 4 undetectable SCE transitions $= \frac{b^4 r}{e^b 4!}\left[\frac{k_{22}(N+1)+1}{N}\right]^4$

Probability of n undetectable SCE transitions $= \frac{b^n r}{e^b n!}\left[\frac{k_{22}(N+1)+1}{N}\right]^n$

Thus, q is given by

$$\mathrm{q} = \frac{b^2 r}{e^b 2!}(K_N)^2 + \frac{b^4 r}{e^b 4!}(K_N)^4 + \cdots$$

$$= \frac{r}{e^b}\left(\frac{(K_N b)^2}{2!} + \frac{(K_N b)^4}{4!} + \cdots\right)$$

$$= \frac{r}{e^b}(\cosh(bK_N) - 1) \tag{15}$$

where $K_N = \left[\frac{k_{22}(N+1)+1}{N}\right]$

Therefore, (14) becomes

$$P_{\tau_B}$$

$$= \left(\frac{b^2 r}{e^b 2!}\frac{2}{N^2}\right)\left(\frac{r}{e^b}\cosh\left(\mathrm{b}\frac{k_{22}(N+1)+1}{N}\right)\right)\left(\frac{r}{e^b} e^{\left(\frac{(k_{11}+k_{13})(N+1)}{N}\right)b}\right) \tag{16}$$

and the ratio of the probability of cells 1 and 2 being sisters (Tree A) *vs.* cousins (Trees B or C) is given by

$$\frac{P_{\tau_A}}{P_{\tau_B}} = \frac{P_{\tau_A}}{P_{\tau_C}} = \frac{P_{\tau_{sisters}}}{P_{\tau_{cousins}}} = \frac{2e^{(1-\frac{(1-k_{22})(N+1)}{N})b}}{\cosh\left(b\frac{k_{22}(N+1)+1}{N}\right)} \tag{17}$$

In this case, the probability ratio is a function of the genomic location of the SCE events, in addition to the SCE rate and the number of bins.

Case III: The second most common and more complicated SCE pattern occurs when an original DNA strand is shared between three cells (Figure 2.2C). Intuitively, Tree B with cells 1 and 3 as sisters is the least likely configuration as it requires one additional SCE transition compared to the other two trees. The probability of observing this SCE pattern in Trees A and C are given by

$$\begin{aligned} P(k_{11}, k_{22}, k_{33}|\tau_A) &= P_{\tau_A} \\ &= \left(\frac{br}{e^b N}\right)\left(\frac{br}{e^b N} e^{b\frac{k_{33}(N+1)}{N}}\right)\left(\frac{r}{e^b} e^{b\frac{(k_{11}+k_{22})(N+1)}{N}}\right) \end{aligned} \tag{18}$$

$$\begin{aligned} P(k_{11}, k_{22}, k_{33}|\tau_C) &= P_{\tau_C} \\ &= \left(\frac{br}{e^b N}\right)\left(\frac{r}{e^b} e^{b\frac{(k_{22}+k_{33})(N+1)}{N}}\right)\left(\frac{br}{e^b N} e^{b\frac{k_{11}(N+1)}{N}}\right) \end{aligned} \tag{19}$$

In (18), the first term accounts for one SCE event between k_{22} and k_{33}. The second term includes one SCE event between $k_{11}$and k_{22} and undetectable SCE events within the right-most genomic region, whose length is equal to k_{33}. The third term accounts for no SCE event within k_{33} and undetectable SCE events within the left region, whose length is equal to $(k_{11} + k_{22})$. Similarly, in (19), the first term accounts for one SCE event between k_{11} and k_{22}. The second term includes no SCE events within k_{11} and undetectable SCE events within the rest of the chromosome, equivalent in length to $(k_{22} + k_{33})$. The third term includes one SCE event between k_{22} and k_{33} and undetectable SCE events within the left-most genomic region. Note that Trees A and C are mirror images of each other and the

probability of observing this SCE pattern is equal for these two tree configurations. For Tree B,

$$P(k_{11}, k_{22}, k_{33}|\tau_B) = P_{\tau_B} = \left(\frac{b^2 r}{e^b N^2}\right)(s)\left(\frac{r}{e^b} e^{b\frac{(k_{11}+k_{33})(N+1)}{N}}\right) \tag{20}$$

The first term is for two SCE events in the first cell division. The second term accounts for an odd number of undetectable SCE transitions within the genomic region between k_{11} and k_{33}, such that both cells 1 and 3 contain parts of the original DNA strand. The third term includes undetectable SCE events within both left and right genomic regions, whose combined length is $(k_{11} + k_{33})$. Further, s is given by

$$\begin{aligned} s &= \frac{\mathrm{b}r}{e^b}(K_N)^1 + \frac{b^3 r}{e^b 3!}(K_N)^3 + \cdots \\ &= \frac{r}{e^b}\left(\frac{(K_N b)^1}{1!} + \frac{(K_N b)^3}{3!} + \cdots\right) \\ &= \frac{r}{e^b}(\sinh(bK_N)) \end{aligned} \tag{21}$$

where $K_N = (\frac{k_{22}(N+1)+1}{N})$.

Therefore, (20) becomes

$$\begin{aligned} &P(k_{11}, k_{22}, k_{33}|\tau_B) = P_{\tau_B} \\ &= \left(\frac{b^2 r}{e^b N^2}\right)\left(\frac{r}{e^b}\sinh\left(\mathrm{b}\frac{k_{22}(N+1)+1}{N}\right)\right)\left(\frac{r}{e^b} e^{b\frac{(k_{11}+k_{33})(N+1)}{N}}\right) \end{aligned} \tag{22}$$

and the ratio of the probability of Tree A *vs.* B is given by

$$\frac{P_{\tau_A}}{P_{\tau_B}} = \frac{P_{\tau_C}}{P_{\tau_B}} = \frac{e^{bk_{22}\frac{N+1}{N}}}{\sinh\left(b\frac{k_{22}(N+1)+1}{N}\right)} \tag{23}$$

Consistent with our intuition, Tree B is less likely than the other two tree topologies, and depending on the values of N, b, and k_{22}, Tree B can be anywhere between 2 to 100 times less likely (Figure 2.2C).

The approach described above can be applied to any SCE pattern. The probability of observing different SCE patterns are estimated for all chromosomes. Next, we assume that each chromosome strand is independent and compute the overall probability of observing the SCE patterns over the whole genome (D) for each Tree i (τ_i). To determine the most likely tree, we compute and compare $P(\tau_A|D)$, $P(\tau_B|D)$, and $P(\tau_C|D)$ using Bayes' theorem

$$P(\tau_i|D) = \frac{P(D|\tau_i) * P(\tau_i)}{P(D)} \tag{24}$$

where $P(\tau_i)$ and $P(D)$ are the probabilities of observing Tree i and the genome-wide SCE pattern data, respectively. $P(\tau_i)$ reflects prior belief of the likelihood that Tree i is the correct topology. As there are 3 possible topologies for any 4-cell tree, we get

$$P(\tau_A|D) + P(\tau_B|D) + P(\tau_C|D) = 1 \tag{25}$$

Further, the ratio of the probability of observing Tree i *vs.* Tree j is given by

$$\frac{P(\tau_i|D)}{P(\tau_j|D)} = \frac{P(D|\tau_i) * P(\tau_i)}{P(D|\tau_j) * P(\tau_j)} = \frac{P(D|\tau_i)}{P(D|\tau_j)} \tag{26}$$

where Tree i *or* j is either Tree *A*, *B*, or *C*. The prior probabilities $P(\tau_i)$ are assumed to be equal to one another, a common practice in Bayesian analysis (Huelsenbeck and Ronquist, 2001). After rearrangement, we get

$$P(\tau_A|D) = \frac{1}{1 + \frac{P(D|\tau_B)}{P(D|\tau_A)} + \frac{P(D|\tau_C)}{P(D|\tau_A)}} \tag{27}$$

Similarly, the probability of all tree topologies can be calculated. Finally, the probability of a particular 8-cell tree is given by the product of the probabilities of the two corresponding 4-cell subtrees.

8. Simulating stand-specific 5hmC distributions

To validate the analytical expressions for the probability of observing different SCE patterns in Figures 2.2B and 2.2C, we simulated 8-cell trees where the occurrence of SCE events were modeled as a Poisson process with $b = 0.3$ and chromosome strands were assumed to segregate randomly ($r = 0.5$). Simulations were performed on chromosome 1 ($N = 97$ for 2 Mb bins). These simulations were then used to estimate the probability of observing Tree A *vs.* Tree B as a function of the position of the SCE event.

To test the accuracy of scPECLR in predicting lineage trees in Figures 2.3B and 2.4A, 8-, 16- or 32-cell embryos with 19 or 38 chromosomes were simulated as described above. All bins in the original DNA strands were hydroxymethylated whereas all subsequently synthesized DNA strands contained no 5hmC, mimicking *in vivo* experimental observations. 5,000 and 2,000 simulated trees were generated for each condition shown in Figures 2.3B and 2.4A, respectively. The trees were subsequently inputted into scPECLR to estimate the percentage of trees that are accurately predicted by the algorithm. For 16-cell trees, we also estimated the prediction accuracy of 2-, 4- and 8-cell subtrees within the full tree, and for 32-cell trees, the 16-cell subtree prediction accuracy was additionally estimated. In the 4-cell embryos in Figure 2.3B, as OSS accurately separates the four cells into two groups of two cells each, the lineage reconstruction problem becomes deterministic, and thus the trees are predicted with 100% accuracy. Similarly, in Figure 2.3C, OSS was assumed to successfully separate the two 8-cell subtrees from 16-cell trees. However, in 32-

cell trees (Figure 2.4A), OSS could not separate cells into two groups in all embryos. Such cases would not continue forward with the calculation and would be classified as incorrect for all level of subtrees. Additionally, the prediction accuracy in 8- and 16-cell trees (Figures 2.3A and 2.3B) were bootstrapped 1000 times. The bootstrap statistics are plotted along with the prediction accuracy.

In the 32-cell tree case where half of the sister pairs are known, 8 out of the 16 sister pairs were randomly selected and became known information about the cells in each simulation. In the 32-cell tree scH&G-seq cases, the genomic variants were also modeled with a Poisson process to occur at a certain rate v per chromosome per cell division. The process starts with the first cell division ($n = 1$, from one to two cells), which has two division actions, generating cells that are the ancestor of cells 1-16 or 17-32. If within a division action, at least one variant emerges, we would assume that we know all cells derived from that particular division action are clustered with one another. For example, if the first division action at $n = 1$ has a variant, we would assume that we know cells 1-16 are clustered together for that simulation. Next, the step proceeds to the two cells dividing at $n = 2$, which has four division actions. Similarly, if the third division action has a variant at $n = 2$, cells 17-24 would be assumed to cluster together. The process continues till $n = 4$, where there are sixteen division actions generating sister pairs. Each cell division action is treated independently. The additional information received from either half the sister pairs or the genomic variants were used to help OSS separate the cells into two groups. If cells are separated into two groups, that simulation trial would continue to the 8-cell grouping step (see "Criteria to determine 32-cell topologies to be evaluated").

9. Consensus tree analysis

This analysis was performed on 16-cell trees to identify parts of the lineage tree that can be predicted with high confidence. The two 8-cell subtrees obtained from OSS are treated independently. The first step is to use a desired relative threshold (RT) to identify all trees that have predicted probabilities within a threshold level of the highest probability tree and include such trees for downstream analysis. All included trees are subsequently weighed equally. The second step is to examine the 4-cell subtrees of each included tree. If all trees consistently predict the same 4-cell subtree, the consensus tree includes the 4-cell subtree. This is true for most datasets as scPECLR largely predicts the 4-cell subtrees accurately in 16-cell trees (Figure 2.3C). When disagreement arises, if the percentage of included trees that have the same 4-cell subtree exceeds a threshold ($t8$), ranging from 0.55 to 1.0, the consensus tree includes the 4-cell subtree, and tree topologies that conflict with this 4-cell subtree are excluded from further analysis. If the percentage is below $t8$, the consensus tree does not include the exact 4-cell subtree but instead attempts to identify as many pairs of cells as possible that appear in different 4-cell subtrees of all included trees, and the consensus analysis terminates. After the 4-cell subtrees are determined, the topology predicted within each of these subtrees is then considered. Again, if all of the remaining trees predict the same topology or if the percentage of remaining trees that predict a consistent topology exceeds a threshold ($t4$), ranging from 0.55 to 1.0, the consensus tree also includes that topology. Otherwise, it does not predict a specific topology within the 4-cell subtree but attempts to identify one cousin pair that appears in the 4-cell topology.

The consensus tree has different levels of specificity, ranging from predicting a full 16-cell tree, where the relationships between all cells are exact, to predicting only two 8-cell

subtrees. In general, each consensus tree is constrained to contain a certain number of tree topologies, which provides information about how specific each consensus tree is. For example, in Figure 2.3D, the consensus tree contains six possible topologies, as there are two topologies arising from uncertainty in the subtree containing cells 5-8 and three topologies arising from uncertainty in the subtree containing cells 13-16. The lower the number of topologies contained within the consensus tree, the more specific and informative it is.

There are three parameters in the consensus tree analysis: RT, $t8$, and $t4$. RT has the largest influence on the structure of the consensus tree, while varying $t8$ and $t4$ leaves the consensus tree largely unchanged (Figures 2.3E-F and Appendix D1: Figure S2.3) (Note: In Figure 2.3E, $t8$ and $t4$ are kept constant at 0.75 and 1, respectively). When the RT increases, the consensus tree becomes more specific but suffers from a higher false discovery rate (FDR). In contrast, although the effects are small, increasing $t8$ and $t4$ leads to a very modest decrease in the specificity of the consensus tree and reduction in FDR. Thus, using different parameter values allows us to tune the competing goals of specificity and accuracy of the consensus tree. In fact, for a specific FDR, there is an optimal set of parameters that gives the most specific consensus tree for a dataset. We performed a consensus tree analysis on the dataset in Figure 2.3B (solid blue lines), with different combinations of RT ranging from 0.05 to 0.50, and $t8$ and $t4$ ranging from 0.55 to 1.0. Each parameter set provides a consensus tree with a different level of specificity, measured by the median number of trees contained in the consensus tree, and the FDR. For any level of FDR tolerated, there is at least one parameter combination that yields the lowest median number of trees. For example, when $b = 0.3$ and the FDR is chosen to be 30%, the optimal

parameter set has RT, $t8$, and $t4$ as 0.05, 0.75, and 1, respectively, yielding the median number of trees contained within the consensus tree to be 36. Thus, for any dataset, the rate of SCE events can be estimated using MLE, and with a user-selected FDR, an optimal parameter set can be estimated to give the most specific consensus tree.

Consensus tree analysis improves the accuracy of lineage prediction in all scenarios. When the SCE rate is low ($b = 0.1$) and the iterative prediction alone performs poorly for 16-cell trees, an error rate of greater than 99% in the iterative prediction decreases to a FDR between 30-75%. When the iterative prediction alone performs moderately ($b = 0.5$), an error rate of ~60% improves to a FDR between 10-45% (Figures 2.3B and 2.3E). Lastly, when the iterative prediction alone performs well ($b = 1.0$), an error rate of ~25% decreases to a FDR between 5-20% (Figures 2.3B and 2.3E). When $b = 1.0$, there are only 1 to 2 median topologies contained in each consensus tree, indicating that the consensus analysis increases the accuracy of the prediction without compromising its specificity. This result shows that scPECLR and the consensus tree analysis provides a significant amount of lineage information with reasonable accuracy for 16-cell trees (Figures 2.3E and 2.3F).

To generate the consensus tree for the 16-cell embryo in Figure 2.3G, 1000 16-cell embryos were simulated with the same SCE rates estimated from the *in vivo* 16-cell embryo. Next, different parameter combinations of RT, $t8$ and $t4$ were used to generate consensus trees. The consensus trees were evaluated against the true tree to calculate FDR rate for each parameter combination. The lowest possible FDR rate of 15% was selected. Subsequently, the parameter combination (RT = 0.05, $t8$ = 0.85, $t4$ = 0.8) that yields the most specific consensus tree with a FDR rate under 15% was chosen for the consensus tree for the *in vivo*

16-cell embryo. There are 180 topologies contained within the consensus tree: 90 from the left 8-cell subtree and 2 from the right 8-cell subtree.

10. Criteria to determine 32-cell topologies to be evaluated

When the cells are successfully separated into two groups of 16 cells, the number of topologies to be considered reduces from more than 10^{26} to ~$4*10^{17}$. We then perform "8-cell grouping", which attempts to further split each 16-cell group into two groups of 8 cells, reducing the number of possible topologies further to fewer than 10^{10}. The first step of 8-cell grouping is to consider all the possible combinations of 16 choose 8 (6435 groupings in total as cells 1-8 grouping and cells 9-16 grouping are considered one grouping). In the case where additional information about the embryos are known, the groupings that conflict with the clonal information were discarded. Next, in each grouping, the 5hmC in all cells within the two 8-cell sets were combined to generate hypothetical 5hmC data of the two cells at the 2-cell stage for that grouping. Then, the number of SCE events present in the hypothetical two cells were calculated. Only the groupings that generate the hypothetical two cells with the fewest number of SCE events were kept. The rationale is that cells accumulate SCE events on their original chromosome strands as they undergo cell division. Therefore, the fewer the SCE events present at the 2-cell stage, the more likely the 8-cell grouping is correct. The left side (cells 1-16) and right side (cells 17-32) undergo the process independently. If there are more than 30 groupings remaining in total, the process is stopped and we would conclude that tree is incorrect for 2-, 4-, and 8-cell subtree levels. The number of remaining groupings, 30, was chosen as there would be about 10000 topologies left after a successful 8-cell grouping. Then scPECLR is used calculate the probabilities of all possible topologies within the four 8-cell sets independently. The topologies that conflict

with known information about the embryo are removed. Then the 8-cell sets for each grouping are combined to generate the full 32-cell tree. The grouping combination that predicts lineage of any cell more than once (i.e. one or more cell is missing from the full tree) is discarded. The probability of the full 32-cell tree is the product of the four probabilities from the 8-cell sets. The full 32-cell tree with the highest probability is the predicted tree.

11. SNP/CNV calling and processing

Variant calling was done via bcftools in a custom shell script. Briefly, the sam file was sorted and only relevant chromosomes (genomic and mitochondria chromosomes) were retained. Next, bcftools called the SNPs with default parameters. SNP calls with quality of at least 20 were kept. A custom Perl script was then used to count the occurrences of SNPs and non-SNPs for each cell at each SNP location in the sam file.

For each library, cells with fewer than 10000 reads were filtered out. 34, 50, and 31 cells passed the cut-off in the AluI, BseRI, and dual enzyme libraries, respectively. Data from autosomes were discretized into 5 Mb bins. Each bin was subsequently normalized by the number of enzyme recognition sites present within that bin. The normalized raw reads were scaled such that the median of the total data is 100. In the dual enzyme library, the normalized raw reads of AluI and BseRI were combined before scaling. The circular binary segmentation (CBS) algorithm was used to call different read count sections in each chromosome. The read count in each bin was then replaced by its mean value from the CBS algorithm. Copy number for each bin was determined by normalizing the mean read count of each cell to two copies and rounding to the nearest integer. To remove outlier bins, the bins that showed more than four copies in any cell were retroactively removed from the

normalized raw read data, which was again inputted into the CBS algorithm. The read count in each bin was once again replaced by its mean value from the CBS algorithm. The copy number of each bin in each cell was subsequently recalculated. The steps were performed independently in each library. All steps of the CBS algorithm have a significance level of 0.01. All cells from the three libraries were combined, with only bins that were present in all libraries retained.

The *clValid* package in R was used to decide between hierarchical, k-means, and pam clustering algorithms and the hierarchical algorithm was recommended. The agglomerative coefficient was then used to determine the appropriate method to calculate distances between cells. Among the options: average, single, complete, and ward, the ward method was recommended. Subsequently, the *NbClust* package was used to determine the number of clusters based on the ward method and Euclidean distance. Two clusters were recommended for the combined data. A dendrogram was created based on the ward method and Euclidean distance.

12. scPECLR is robust to initial estimates of the SCE rate and to varying SCE rates at each cell division

We explored the robustness of scPECLR to initial estimates of the SCE rate by simulating strand-specific 5hmC data in 8-cell trees with a constant SCE rate ($b = 0.3$). We then used different values of SCE rates – ranging from 0.1 to 2.0 – in scPECLR to predict the lineage tree (instead of estimating the SCE rate from the observed SCE pattern using MLE). We found that the percentage of trees that were accurately predicted did not change over the range of SCE rates, suggesting that scPECLR is robust to uncertainty in SCE rate

estimation and the prediction accuracy mainly depends on the SCE rates used to generate the 5hmC data (Appendix D1: Figures S2.5A and S5B).

As the 8-cell mouse embryos have varying rates of SCE events across cell divisions, we explored the robustness of scPECLR when the rates are different for each cell division. Because prediction accuracy of scPECLR is dependent on the rate of SCE events, in this analysis, we fixed the combined SCE rate (B) over 3 (or 4) cell divisions, but allowed individual cell divisions to have different rates. For 8-cell trees, the model is largely robust against varying rates of SCE events across cell divisions, with higher B and larger number of chromosomes resulting in better prediction accuracy (Appendix D1: Figure S2.5C). For example, when the SCE rates are low for the first and second cell division (b_1 and b_2) and high for the third cell division (b_3), similar to the experimental observation in 8-cell mouse embryos, scPECLR predicts the lineage tree with very high accuracy (Appendix D1: Figure S2.5C, H3). One case where the prediction accuracy drops modestly is when the SCE rates of the first and third cell divisions (b_1 and b_3) are low and the SCE rate of the second cell division (b_2) is high (Appendix D1: Figure S2.5C, H2). In this case, the data has a large number of SCE events that are shared between cousin cells. As the SCE rate at each cell division is assumed constant during the first iteration of scPECLR, the algorithm predicts that cells sharing more SCE events are more likely to be sisters. This misidentification results in a large percentage of simulations not predicting the true tree after the first iteration. However, the prediction improves significantly after a few iterations because starting from the second iteration, the model accounts for different SCE rates at each cell division. Consequently, the varying SCE rates at each cell division has minimal impact on the accuracy of 8-cell tree prediction.

For 16-cell trees, there are a few cases where the prediction accuracy is worse than when the rates are uniformly distributed; these include situations where b_4 is low (Appendix D1: Figure S2.5D, H2, H3, H13, H23, and L4). In these cases, the prediction accuracy is lower because scPECLR inaccurately infers a pair of cousin or second cousin cells as sister cells due to a large number of SCE events shared between such pairs. In contrast, cases with high b_4 values result in better prediction accuracy because scPECLR correctly identifies sister cell pairs (Appendix D1: Figure S2.5D, H4, H14, H24, and H34). Finally, scPECLR also performs well when b_2 and b_3 are low as it does not misidentify cousin or second cousin pairs as sister pairs. These results suggest that in addition to the combined SCE rate, how the individual SCE rates are distributed over each cell division impacts the accuracy of reconstructing 16-cell trees.

13. Statistical test to identify non-random DNA segregation

To test the segregation pattern of DNA strands at the 4-cell stage, the 5hmC profile of 8-cell mouse embryos were combined using the lineages predicted by scPECLR to obtain the distribution of 5hmC on the original DNA strands at the 4-cell stage, while the *in vivo* experimental 4-cell mouse embryo data could be used without prior processing. If a majority of an original chromosome strand is present in one cell at the 4-cell stage, that cell is considered to inherit the entire chromosome strand. This is to account for the limited number of original strands that undergo a few SCE events during cell division. A binomial two-tailed test was conducted with a null hypothesis of random segregation ($\pi = 0.5$) and an alternative hypothesis of non-random segregation ($\pi \neq 0.5$). Two pairs of sister cells from 27 embryos were considered to display statistically significant non-random DNA segregation

for p-values lower than 0.05, one pair from the 4-cell embryo dataset and the other from the 8-cell embryo dataset.

To test whether the two events of non-random segregation can be explained by chance alone, we randomly sampled 27 embryos from a pool of 100000 simulated 4-cell randomly-segregating embryos, generated with a constant SCE rate of $b = 0.3$, and counted how many events of non-random segregation with $p < 0.05$ were found. The random sampling was conducted 10000 times. The cumulative distribution of the number of non-random segregation events found was plotted in Figure 2.5D. Despite a median of one event, we failed to reject the null hypothesis that two events of non-random segregation could be explained by chance alone.

B. *Chapter 3 Methods*

1. Bacterial strains and culture conditions

Escherichia coli MG1655 (ATCC: 700926) overnight cultures were inoculated into fresh LB medium at 1:50 and grown at 37°C with shaking (150 rpm). Upon reaching the exponential growth phase, the culture was centrifuged at 3000 g for 10 min. The media was removed and the pellet was resuspended in PBS to a concentration of 10^7 cells per µL. The cells were stored on ice and total RNA extraction was performed immediately.

2. RNA extraction

Trizol (Thermo Fisher Scientific, Cat. # 15596018) RNA extraction was performed following the manufacturer's protocol. Briefly, 10^8 cells were added to 750 µL Trizol, mixed, and then combined with 150 µL chloroform. After centrifugation, the clear aqueous layer was recovered and precipitated with 375 µL of isopropanol and 0.67 µL of GlycoBlue

(Thermo Fisher Scientific, Cat. # AM9515). The pellet was washed twice with 75% ethanol and after the final centrifugation, the resulting pellet was resuspended in RNase-free water.

3. EMBR-seq

Poly adenylation. 100 ng of total RNA in 2 µL was combined with 3 µL poly(A) mix, comprised of 1 µL 5x first strand buffer [250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$, comes with Superscript II reverse transcriptase, Invitrogen Cat. # 18064-014], 1 µL blocking primer mix (see Appendix B8: *Primers*), 0.8 µL nuclease-free water, 0.1 µL 10 mM ATP, and 0.1 µL *E. coli* poly(A) polymerase (New England Biolabs, Cat. # M0276S). The mixture was incubated at 37°C for 10 min. In the control group, no blocking primers were added and 1.8 µL of nuclease-free water was added instead. For EMBR-seq with either unmodified or phosphorylated 3'-end blocking primers, the blocking primer mix was prepared by mixing equal volumes of 50 µM blocking primers specific to 5S, 16S and 23S rRNA. For EMBR-seq with hotspot blocking primers, the blocking primer mix was prepared by mixing equal volumes of 100 µM 3'-end blocking primers with 100 µM hotspot blocking primers, such that the final mixture was 50 µM 3'-end primers (3 primers mixed) and 50 µM hotspot primers (6 primers mixed).

Reverse transcription. The polyadenylation product was mixed with 0.5 µL 10 mM dNTPs (New England Biolabs, Cat. # N0447L), 1 µL reverse transcription primers (25 ng/µL, see Appendix B8: *Primers*), and 1.3 µL blocking primer mix, and heated to 65°C for 5 min, 58°C for 1 min, and then quenched on ice. In the control samples, the blocking primers were again replaced with nuclease-free water. Next, 3.2 µL RT mix, consisting of 1.2 µL 5x first strand buffer, 1 µL 0.1 M DTT, 0.5 µL RNaseOUT (Thermo Fisher Scientific, Cat. #10777019), and 0.5 µL Superscript II reverse transcriptase was added to the

solution, followed by 1 h incubation at 42°C. The temperature was then raised to 70°C for 10 min to heat inactivate Superscript II.

Second strand synthesis. 49 µL of the second strand mix, containing 33.5 µL water, 12 µL 5x second strand buffer [100 mM Tris-HCl (pH 6.9), 23 mM $MgCl_2$, 450 mM KCl , 0.75 mM β-NAD, 50 mM $(NH_4)_2$ SO_4, Invitrogen, Cat. # 10812-014], 1.2 µL 10 mM dNTPs, 0.4 µL *E. coli* ligase (Invitrogen, Cat. # 18052-019), 1.5 µL DNA polymerase I (Invitrogen, Cat. # 18010-025), and 0.4 µL RNase H (Invitrogen, Cat. # 18021-071), was added to the product from the previous step. The mixture was incubated at 16°C for 2 h. cDNA was purified with 1x AMPure XP DNA beads (Beckman Coulter, Cat. # A63881) and eluted in 24µL nuclease-free water that was subsequently concentrated to 6.4 µL.

***In vitro* transcription.** The concentrated solution was mixed with 9.6 µL of Ambion *in vitro* transcription mix (1.6 µL of each ribonucleotide, 1.6 µL 10x T7 reaction buffer, 1.6 µL T7 enzyme mix, MEGAscript T7 Transcription Kit, Thermo Fisher Scientific, Cat. # AMB13345) and incubated at 37°C for 13 h. Next, the aRNA was treated with 6 µL EXO-SAP (ExoSAP-IT™ PCR Product Cleanup Reagent, Thermo Fisher Scientific, Cat. # 78200.200.UL) at 37°C for 15 min followed by fragmentation with 5.5 µL fragmentation buffer (200 mM Tris-acetate (pH 8.1), 500 mM KOAc, 150 mM MgOAc) at 94°C for 3 min. The reaction was then quenched with 2.75 µL stop buffer (0.5 M EDTA) on ice. The fragmented aRNA was size selected with 0.8x AMPure RNA beads (RNAClean XP Kit, Beckman Coulter, Cat. # A63987) and eluted in 15 µL nuclease-free water. Thereafter, Illumina libraries were prepared as described previously (Hashimshony et al., 2016).

4. EMBR-seq with TEX digestion

To test the Terminator™ 5'-phosphate-dependent exonuclease (Lucigen, Cat. # TER5120), 100 ng of total RNA in 2 µL was combined with 18 µL TEX mix, comprised of 14.5 µL nuclease free water, 2 µL Terminator 10x buffer A, 0.5 µL RNAseOUT, and 1 µL TEX. The solution was incubated at 30°C for 1 h and quenched with 1 µL of 100 mM EDTA. The product was purified with 1x AMPure RNA beads and eluted in 10 µL nuclease-free water and concentrated to 2 µL. This TEX digested total RNA was then used as starting RNA in the EMBR-seq protocol described above.

5. EMBR-seq bioinformatic analysis

Paired-end sequencing of the EMBR-seq libraries was performed on an Illumina NextSeq 500. All sequencing data has been deposited to Gene Expression Omnibus under the accession number GSE149666. In the sequencing libraries, the left mate contains information about the sample barcode (see Appendix B8: *Primers*). The right mate is mapped to the bacterial transcriptome. Prior to mapping, only reads containing valid sample barcodes were retained. Subsequently, the reads were mapped to the reference transcriptome (*E. coli* K12 substr. MG1655 cds ASM584v2) using Burrows-Wheeler Aligner (BWA) with default parameters.

6. Analysis of detection bias in EMBR-seq

E. coli operons were downloaded from RegulonDB (Santos-Zavaleta et al., 2019). Operons with at least 2 genes were included for this analysis. The data from EMBR-seq libraries with 100 ng starting material was mapped to *E. coli* K12 substr. MG1655 reference genome (ASM584v2). For each read that maps within an operon, the distance of the mapped

location from the 3’ end of the operon was calculated, accounting for the read length. Next, the operons were discretized into 50 bins, and all operons with more than 200 unique reads were considered for downstream analysis. The number of reads in each bin was then normalized by the total number of reads in each operon, and the average of the relative reads within each bin was calculated. To compare bacterial data from EMBR-seq to mammalian data from CEL-seq, we downloaded CEL-seq data reported in Grün *et al.* (GEO Accession: GSM1322290) and performed similar analysis for the mouse genes (Grün et al., 2014).

7. Sequence conservation of 16S and 23S rRNA

16S rRNA sequences from 4000 species were obtained from *rrnDB* , while 23S rRNA sequences from 119 species were selected from NCBI RefSeq (Stoddard et al., 2015). Next, the last 100 bases from the 3’end of each sequence were aligned using Clustal Omega (O’Leary et al., 2016). Shannon entropy for each aligned base location was then calculated such that the maximal entropy value was 1. Five possibilities were allowed: “A”, “T”, “C”, “G”, and “-” (Madeira et al., 2019).

8. Primers

Reverse transcription primers are shown below with the 6-nucleotide sample barcodes underlined (Hashimshony et al., 2016):

GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNN(NNNNNN)TTTTTTTTTTTTTTTTTTTTTTTTV

The following five barcodes were used in this study:

AGACTC

AGCTTC

CATGAG

CAGATC

TCACAG

Blocking primers:

5S 5’-ATGCCTGGCAGTTCCCTACTCTCGCATGGG-3’

16S 5’-TAAGGAGGTGATCCAACCGCAGGTTCCCCT-3’

23S 5’-AAGGTTAAGCCTCACGGTTCATTAGTACCG-3’

In the case of the 3’ phosphorylated primers, all blocking primers have a 3’ phosphorylation modification.

Hotspot blocking primers:

16S primer for hotspot at position 107:

5’-GGCACATCCGATGGCAAGAGGCCCGAAGGT-3’

16S primer for hotspot at position 682:

5’-TCCTGTTTGCTCCCCACGCTTTCGCACCTG-3’

16S primer for hotspot at position 1241:

5’-CCGTGGCATTCTGATCCACGATTACTAGCGATTCCG-3

23S primer for hotspot at position 375:

5’-CGCCTTTCCCTCACGGTACTGGTTCACTATCGG-3’

23S primer for hotspot at position 1421:

5’-TTGCTTCAGCACCGTAGTGCCTCGTCATCA-3’

23S primer for hotspot at position 1641:

5’-GCAGCCAGCTGGTATCTTCGACTGATTTCAGC-3’

Each primer is designed to anneal approximately 100 bp downstream of the hotspot. The exact position and length of each primer was adjusted to ensure the T_m was above 65°C.

C. *Chapter 4 Methods*

1. EMBR-RNase H (EMBR-H)

The amplified RNA (aRNA) was made as described in Appendix B3. First, 6 µL of ExoSAP-IT enzyme was added to 16 µL of sample, followed by 15 minutes incubation at 37 ^{o}C. Next, 5.5 µL of fragmentation buffer was added, followed by 3 minutes incubation at 94 ^{o}C, immediately followed by the addition of 2.75 µL fragmentation stop buffer. Next, 0.8x RNA cleanup was performed, with 20 µL elution. aRNA concentration is then measured and the amount of aRNA in ng (9 µL) is calculated.

For Thermostable RNase H:

To 9 µL of aRNA, add the following mix (0.5x mass amount of hybridization primers, 3.2 µL of 5X first-strand buffer, and water to total of 15 µL). Incubate 65 ^{o}C for 2 minutes and quench samples on ice. Add 1 µL of Hybridase RNase H or water (in case of control) was added, followed by 30 minutes incubation at 45 ^{o}C (lid at 60 ^{o}C). After the reaction, immediately place the samples on ice.

For standard RNase H:

First, mix 10 µL of 5X first-strand buffer with 2.8 µL of 5X second strand buffer to create 5X FS/SS buffer mix. To 9 µL of aRNA, add the following mix (0.5x mass amount of hybridization primers, 3.2 µL of 5X FS/SS buffer mix, and water to total of 15 µL). Incubate at 65 ^{o}C for 2 minutes and quench samples on ice. Add 1 µL of Hybridase RNase H or water

(in case of control) was added, followed by 30 minutes incubation at 16 ºC (lid at 40 ºC). After the reaction, immediately place the samples on ice.

For all samples, 1X RNA bead cleanup with 15 µL was performed, followed by vacufuge to 5 µL. Concentrated sample solutions then undergo RT and PCR in a similar protocol described in EMBR-seq

2. rRNA hybridization primers

rRNA hybridization primers were designed to target the 3' end and hotspot locations of rRNA. Following the design in Huang *et al*., the primer lengths were extended to 50 nucleotides (Huang et al., 2020). Note that since amplified RNA, which is complementary to the total RNA, is the target of RNase H, rRNA hybridization primers also target the complementary strand. Generally, rRNA hybridization primers are the reverse complement of the blocking primers, extended to 50 bp to the 5' side of the probe. Since 5S rRNA was rarely detected in EMBR results. No primers were designed to deplete 5S rRNA.

16S primer for 3' end:

5'- AGT CGT AAC AAG GTA ACC GTA GGG GAA CCT GCG GTT GGA TCA CCT CCT TA -3'

16S primer for hotspot at position 107:

5'- TCG CAA GAC CAA AGA GGG GGA CCT TCG GGC CTC TTG CCA TCG GAT GTG CC -3'

16S primer for hotspot at position 682:

5'- CAG GTG CGA AAG CGT GGG GAG CAA ACA GGA TTA GAT ACC CTG GTA GTC CA -3'

16S primer for hotspot at position 1241:

5’- CGA CTC CAT GAA GTC GGA ATC GCT AGT AAT CGT GGA TCA GAA TGC CAC GG -3’

23S primer for 3’ end:

5’- CAG CGA TGC GTT GAG CTA ACC GGT ACT AAT GAA CCG TGA GGC TTA ACC TT - 3’

23S primer for hotspot at position 375:

5’- CCG ATA GTG AAC CAG TAC CGT GAG GGA AAG GCG AAA AGA ACC CCG GCG AG -3’

23S primer for hotspot at position 1421:

5’- GGA AAA TCA AGG CTG AGG CGT GAT GAC GAG GCA CTA CGG TGC TGA AGC AA -3’

23S primer for hotspot at position 1641:

5’- AAG CGA CTT GCT CGT GGA GCT GAA ATC AGT CGA AGA TAC CAG CTG GCT GC -3’

3. EMBR-TtAgo (EMBR-T)

There are 6 different conditions: the non-EMBR and EMBR cases (3 each) and the noTtAgo cases: cDNA product as single-stranded DNA target of TtAgo and PCR product as double-stranded DNA target of TtAgo (2 each).

Follow the same steps as described in EMBR (with or without blocking primers) until library preparation step.

RT reaction:

For the noTtAgo cases, start with 5 µL of aRNA, add 1 µl randomhexRT primer of cel-seq2 and 0.5 µL dNTPs, followed by 5 minutes incubation at 65 ^{o}C. Place the samples on

ice. Add 4 μL of the solution mix, containing 2 μL first strand buffer, 1 μL 0.1 M DTT, 0.5 μL RNaseOUT, and 0.5 μL Superscript II, to the total of 10.5 μL.

For TtAgo case, start with ~13 μL of aRNA, add 1 μl randomhexRT primer of cel-seq2 and 0.5 μL dNTPs, followed by 5 minutes incubation at 65 ºC. Place the samples on ice. Add 6 μL of the solution mix, containing 4 μL first strand buffer, 1 μL 0.1 M DTT, 0.5 μL RNaseOUT, and 0.5 μL Superscript II, to the total of 20.5 μL.

For all samples, incubate at 25 ºC for 10 minutes, 42 ºC for 1 hour, and 70 ºC for 10 minutes.

TtAgo:

For the noTtAgo case, skip this step.

For single-stranded DNA target cases (without pre-PCR), split the aRNA-RT solution in half (10.25 μL each). Save one half in -80 ºC. To the other half, perform 1x DNA bead cleanup, eluted in 15 μL of water. Next, add 2 μL of 10X ThermoPol Buffer and 1 μL of 5 μM ssDNA (-). Add water until the total volume is 19 μL. Incubate at 95 ºC for 5 minutes and reduce the temperature to 80 ºC. Then add 1 μL of TtAgo.

For double-stranded DNA target cases (with pre-PCR), split the aRNA-RT solution in half (10.25 μL each). Save one half in -80 ºC. To the other half, add 0.25 μL of water, 12.5 μL PCR mix, 1 μL RP1, and 1 μL of uniquely indexed PCR primer. Perform 3 PCR cycles. Perform 1x DNA bead cleanup, eluted in 15 μL water. Next, add 2 μL of 10X ThermoPol Buffer, 1 μL of 5 μM ssDNA (-) and 1 μL of 5 μM ssDNA (+). Add water until the total volume is 19 μL. Then add 1 μL of TtAgo.

For all TtAgo cases, incubate at 80 ºC for 30 minutes, followed by rapid cooling to 4 ºC. Perform 1x DNA bead cleanup, eluted in 15 μL water. Vacufuge to condense the solution

down to 5.25 μL. Then proceed to finish library preparation at the PCR step described in EMBR-seq. Note that 3 more PCR cycles are performed for TtAgo cases without pre-PCR and the same uniquely indexed PCR primers for each TtAgo cases with pre-PCR were used.

4. TtAgo guide primers

The following guidelines were taken into consideration when designing guide primers. The guide primers all have 5' phosphate and should be 16-18 nucleotides in length (all primers designed are 16 bp long). The guide primers must start with a T and do not have an A in the 12th position. For the protocol with single-stranded cDNA target, only antisense (-) guide primers were added, while both sense (+) and antisense (-) guide primers are used when the target is double-stranded. To prevent the primers from annealing to each other, their target locations are slightly staggered. Since 5S rRNA was rarely detected in EMBR results, no primers were designed to deplete 5S rRNA.

16S primer for 3' end:

(-) 5'- TAAGGAGGTGATCCAA -3'

(+) 5'- TAACAAGGTAACCGTA -3'

16S primer for hotspot at position 107:

(-) 5'- TACTAGCTAATCCCAT -3'

(+) 5'- TAACGGCTCACCTAGG -3'

16S primer for hotspot at position 682:

(-) 5'- TATCTAATCCTGTTTG -3'

(+) 5'- TAGTCCACGCCGTAAA -3'

16S primer for hotspot at position 1241:

(-) 5'- TTCTGATCCACGATTA -3'

(+) 5’- TGAAGTCGGAATCGCT -3’

23S primer for 3’ end:

(-) 5’- TTCATTAGTACCGGTT -3’

(+) 5’- TTGAAGACGACGACGT -3’

23S primer for hotspot at position 375:

(-) 5’- TTTCCCTCACGGTACT -3’

(+) 5’- TGACCGATAGTGAACC -3’

23S primer for hotspot at position 1421:

(-) 5’- TTGCTTCAGCACCGTA -3’

(+) 5’- TCTAAGCATCAGGTAA -3’

23S primer for hotspot at position 1641:

(-) 5’- TATCTTCGACTGATTT -3’

(+) 5’- TGGCTCTGTTTATTAA -3’

5. EMBR-seq in mammalian cells

There are 6 different conditions: the non-EMBR and EMBR cases (3 each) and the cleanup cases, where after the PCR step of the library preparation, either one round of right-side cleanup or two rounds of cleanups is performed (2 each). Mouse and human cells undergo the same protocol with their appropriate blocking primers. Follow the same steps as described in EMBR-seq. For the conditions that have two rounds of bead cleanups, follow the bead cleanup described in EMBR.

6. Mammalian blocking primers

Mouse cells

5S primer for 3’ end:

5’- AAAGCCTACAGCACCCGGTATTCCCAGGCG -3’

5.8S primer for 3’ end:

5’- CAACCGACGCTCAGACAGGCGTAGCCC -3’

18S primer for 3’ end:

5’- TTAATGATCCTTCCGCAGGTTCACCTACGGAAACC -3’

18S primer for hotspot at position 122:

5’- GGAGGGAGCTCACCGGGTTGGTTTTGATCT -3’

18S primer for hotspot at position 636:

5’- TAAGAGCATCGAGGGGGCGCCGAGAGGCAA -3’

18S primer for hotspot at position 1318:

5’- TTGTCCCTCTAAGAAGTTGGGGGACGCCGA -3’

28S primer for 3’ end:

5’- GAAAGCCCGCAGAGACAAACCCTTGTGTCG -3’

28S primer for hotspot at position 324:

5’- ACTGCGCGGACCCCACCCGTTTACCTCTTAA -3’

28S primer for hotspot at position 1257:

5’- TACGGACCTCCACCAGAGTTTCCTCTGGCTTCG -3’

28S primer for hotspot at position 2145:

5’- ATTCGGGGATCTGAACCCGACTCCCTTTCGAT -3’

28S primer for hotspot at position 3551:

5’- ACGATGAGAGTAGTGGTATTTCACCGGCGGC -3’

28S primer for hotspot at position 4644: This primer targeted close to the 3’ end of 28S was added with VV overhang as the 3’ end 28S primer did not appear to be very effective.

5’- VVGAAAGCCCGCAGAGACAAACCCTTGTGTCG -3’

Human cells

5S primer for 3’ end: same as mouse

5.8S primer for 3’ end:

5’- AAGCGACGCTCAGACAGGCGTAGCCCC -3’

18S primer for 3’ end:

5’- TAATGATCCTTCCGCAGGTTCACCTACGGAAACC -3’

18S primer for hotspot at position 122:

5’- CGGCCCGAGGTTATCTAGAGTCACCAAAGC -3’

18S primer for hotspot at position 635: same as mouse 18S, position 636

18S primer for hotspot at position 1318: same as mouse 18S, position 1318

28S primer for 3’ end:

5’- GACAAACCCTTGTGTCGAGGGCTGACTTTCAATAG -3’

28S primer for hotspot at position 325: same as mouse 28S, position 324

28S primer for hotspot at position 983:

5’- ACGCGCGCGTGGCCCCGAGAGAACCT -3’

28S primer for hotspot at position 1968:

5’- TTCAAGGCTCACCGCAGCGGCCCTCCTACT -3’

28S primer for hotspot at position 2390: same as mouse 28S, position 2145

28S primer for hotspot at position 4656:

5’- ACCGGCTATCCGAGGCCAACCGAGGCT -3’

7. Mammalian EMBR-seq data processing pipeline

The following data processing pipeline is the same for mouse and human cells. The sequencing reads were trimmed to the desired length: 25 bp for R1 and 30 bp for R2, unless otherwise noted. TrimGalore was used to remove any remaining Cel-Seq barcodes present (Martin, 2011). Reads with Cel-Seq barcodes were extracted and mapped to the constructed reference file. ERCC reads were removed and an optional down sample step was performed to a desired read depth. Each mapped reads was categorized to mRNA, rRNA, or ncRNA. Each reads in ncRNA was further classified to each subtype: sRNA, scRNA, scaRNA, snRNA, snoRNA, miRNA, lncRNA, and lincRNA.

Reference file was constructed by concatenating mRNA, rRNA and ncRNA files together. mRNA transcriptome model was chosen from RefSeq (O'Leary et al., 2016); rRNA sequences were from NCBI (Stoddard et al., 2015), and ncRNA file was made from GENCODE annotation (Frankish et al., 2019). In the case of mouse, there are some overlapped genes between RefSeq transcriptome model and GENCODE annotation. MGI annotation was used to resolve conflicting annotation (Bult et al., 2019).

D. *Supplementary Figures*

1. Chapter 2

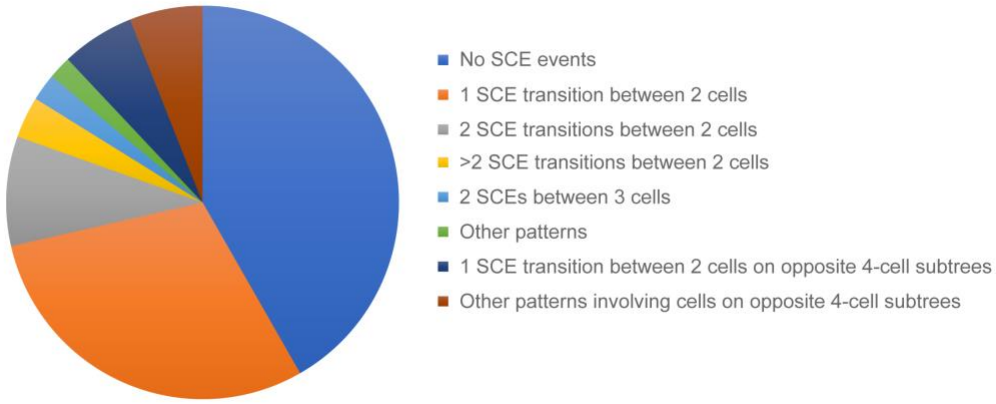


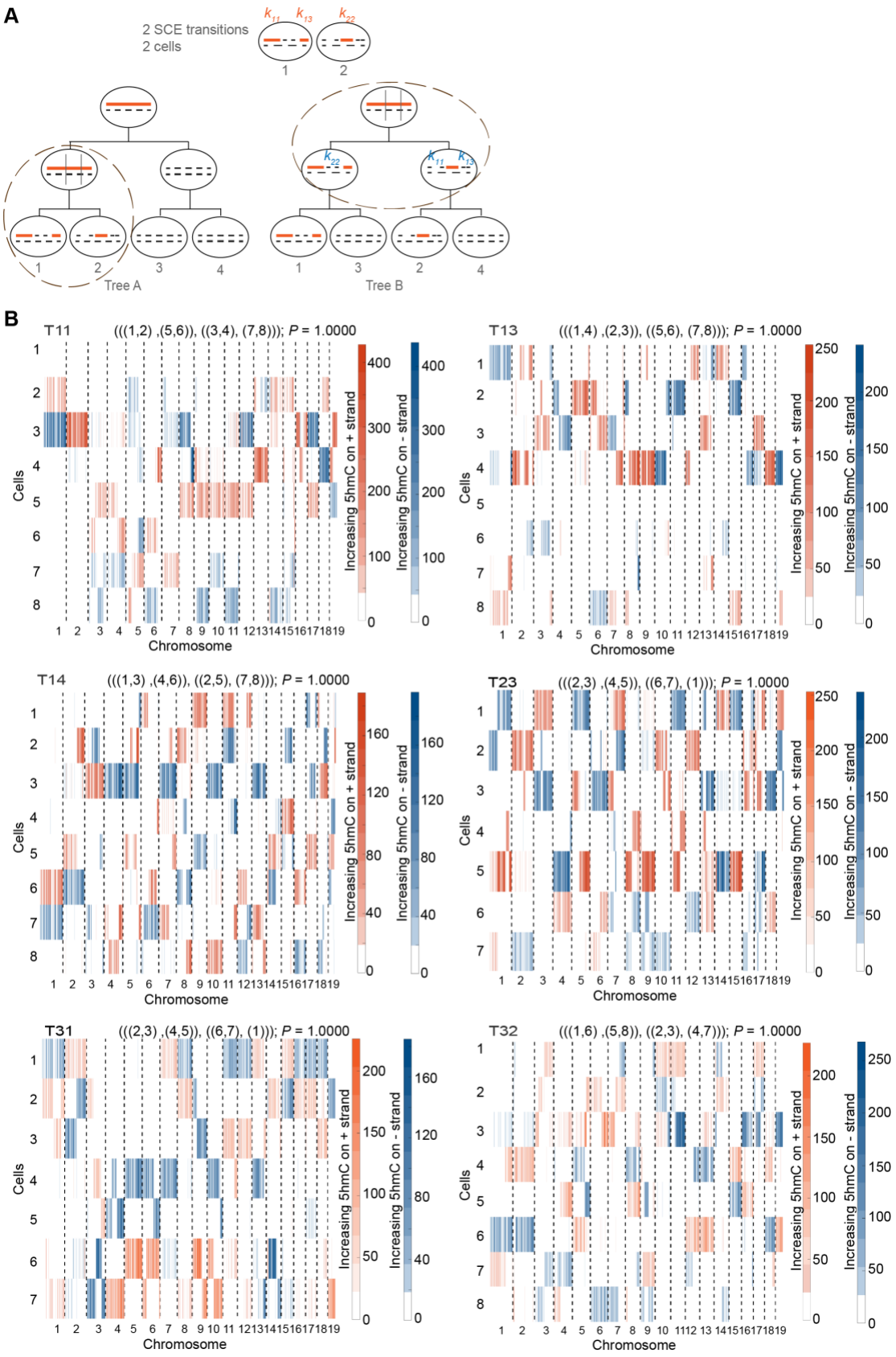
Figure S 1.1 Distribution of SCE patterns in 8-cell mouse embryos

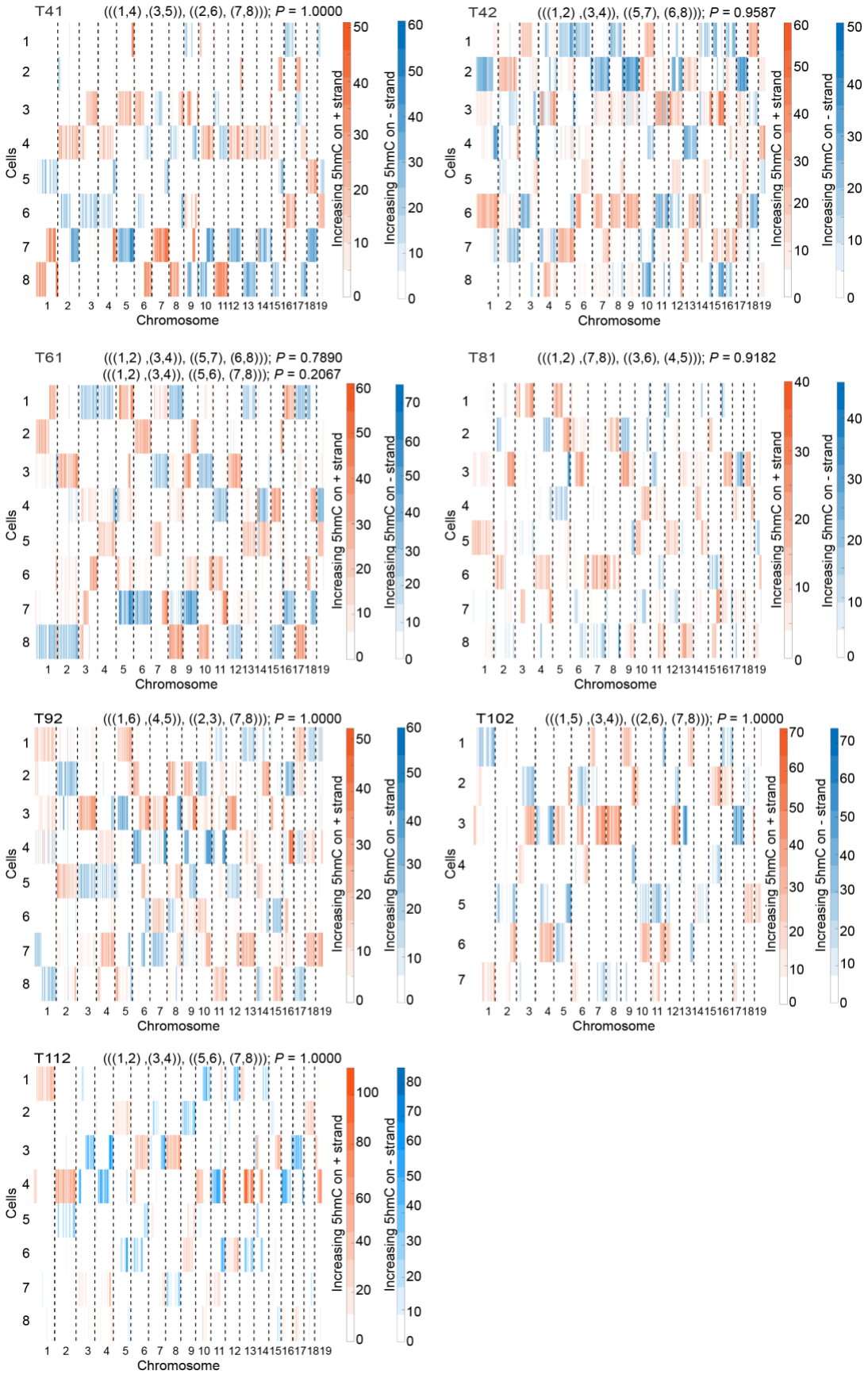
Approximately 33% of the original DNA strands display one SCE transition that is shared between two cells within the same 4-cell subtree (orange). In addition to this most frequently observed pattern, a large diversity of other SCE patterns are observed in 8-cell mouse embryos. All observed SCE patterns are used to probabilistically reconstruct cellular lineages in scPECLR. Approximately 40% of the original paternal DNA strands do not undergo SCE events in the first three cell divisions up to the 8-cell stage of mouse embryogenesis.

Figure S 1.2 Reconstructing lineage trees for preimplantation mouse embryos using scPECLR

(A) The more complex pattern of two SCE transitions shared between two cells increasingly favors the sister tree to the cousin tree arrangement. Schematic showing the SCE events that are necessary for two cells that share two SCE transitions to be sister (Tree A) or cousin cells (Tree B). Mathematically, the original DNA strand undergoes the same number of SCE transitions in both tree topologies and the probability of observing the SCE event shown within the dotted circles is identical for Trees A and B. Further, in Tree A, the cell division that gives rise to cells 3 and 4 is unconstrained in the number of SCE events that can take place. In contrast, while any number of SCE events can occur within the k_{11} and k_{13} genomic regions in Tree B, the k_{22} region is constrained to have an even number of SCE events, thereby reducing the likelihood to observing Tree B compared to Tree A.

(B) The genome-wide strand-specific 5hmC distribution of ten 8-cell and three 7-cell mouse embryos are shown. The 7-cell mouse embryos contain one blastomere from the 4-cell stage of embryogenesis that had not yet divided at the time the embryos were isolated. The mosaic pattern of 5hmC and the SCE events can be used to reconstruct the cellular lineages using scPECLR. The predicted lineage tree and the probability of observing this topology is indicated above each panel. Reconstructing the 7-cell embryos T23, T31, and T102 shows that scPECLR can be applied to non-symmetric trees. Note that for two 8-cell mouse embryos, we were not able to successfully sequence the 5hmC of one cell in each embryo (cell 1 in the embryo T11 and cell 5 in the embryo T13). By assuming that the original DNA strands that were not observed in any of the remaining 7 cells must have been present in the cell that failed to sequence, we were able to successfully predict the 8-cell lineage tree. These results suggest that scPECLR can also be used in cases where there is a limited amount of missing 5hmC sequencing data.


A
2 SCE transitions
2 cells
k_{11} k_{13} k_{22}
1 2
Tree A
Tree B
B
T11 (((1,2) ,(5,6)), ((3,4), (7,8))); P = 1.0000
T13 (((1,4) ,(2,3)), ((5,6), (7,8))); P = 1.0000
T14 (((1,3) ,(4,6)), ((2,5), (7,8))); P = 1.0000
T23 (((2,3) ,(4,5)), ((6,7), (1))); P = 1.0000
T31 (((2,3) ,(4,5)), ((6,7), (1))); P = 1.0000
T32 (((1,6) ,(5,8)), ((2,3), (4,7))); P = 1.0000
Cells
Chromosome
Increasing 5hmC on + strand
Increasing 5hmC on - strand


T41
(((1,4) ,(3,5)), ((2,6), (7,8))); P = 1.0000
T42
(((1,2) ,(3,4)), ((5,7), (6,8))); P = 0.9587
T61
(((1,2) ,(3,4)), ((5,7), (6,8))); P = 0.7890
(((1,2) ,(3,4)), ((5,6), (7,8))); P = 0.2067
T81
(((1,2) ,(7,8)), ((3,6), (4,5))); P = 0.9182
T92
(((1,6) ,(4,5)), ((2,3), (7,8))); P = 1.0000
T102
(((1,5) ,(3,4)), ((2,6), (7,8))); P = 1.0000
T112
(((1,2) ,(3,4)), ((5,6), (7,8))); P = 1.0000
Cells
Chromosome
Increasing 5hmC on + strand
Increasing 5hmC on - strand

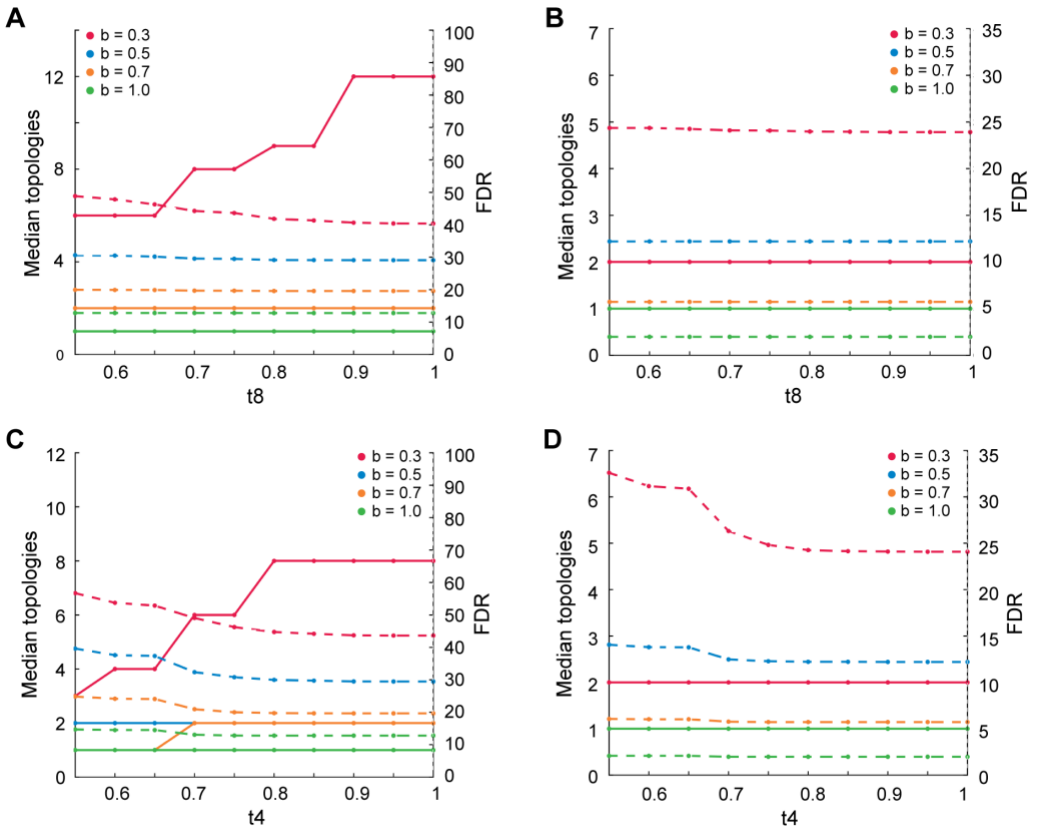


Figure S 1.3 Parameters $t8$ and $t4$ have minor impact on the consensus tree analysis

Panels show representative examples of how the median number of topologies of the consensus tree and the FDR varies with $t8$ and $t4$. These plots are shown for (A) $RT = 0.25$ and $t4 = 1$ for 16-cell trees with 19 chromosomes; (B) $RT = 0.25$ and $t4 = 1$ for 16-cell trees with 38 chromosomes; (C) $RT = 0.25$ and $t8 = 0.75$ for 16-cell trees with 19 chromosomes; and (D) $RT = 0.25$ and $t8 = 0.75$ for 16-cell trees with 38 chromosomes. Solid lines indicate the median number of topologies contained in the consensus tree on the left axis, and the dotted lines indicate the FDR on the right dotted axis. Varying $t8$ and $t4$ across the entire range of values shows that it does not have a significant impact on the median number of topologies contained in the consensus tree or the FDR. Note that in panel (A), the blue solid line is covered by the yellow solid line as they have the same number of median topologies in all cases. Similarly, in panels (B) and (D), both the blue and yellow solid lines are covered by the green solid line.

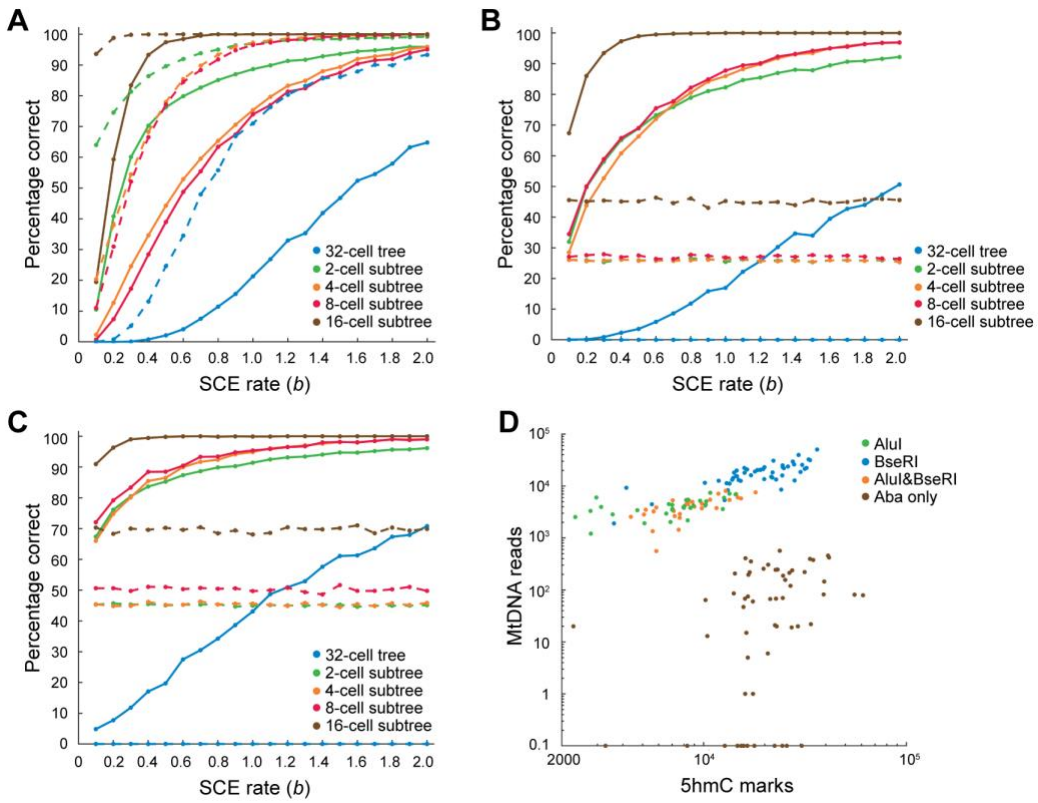
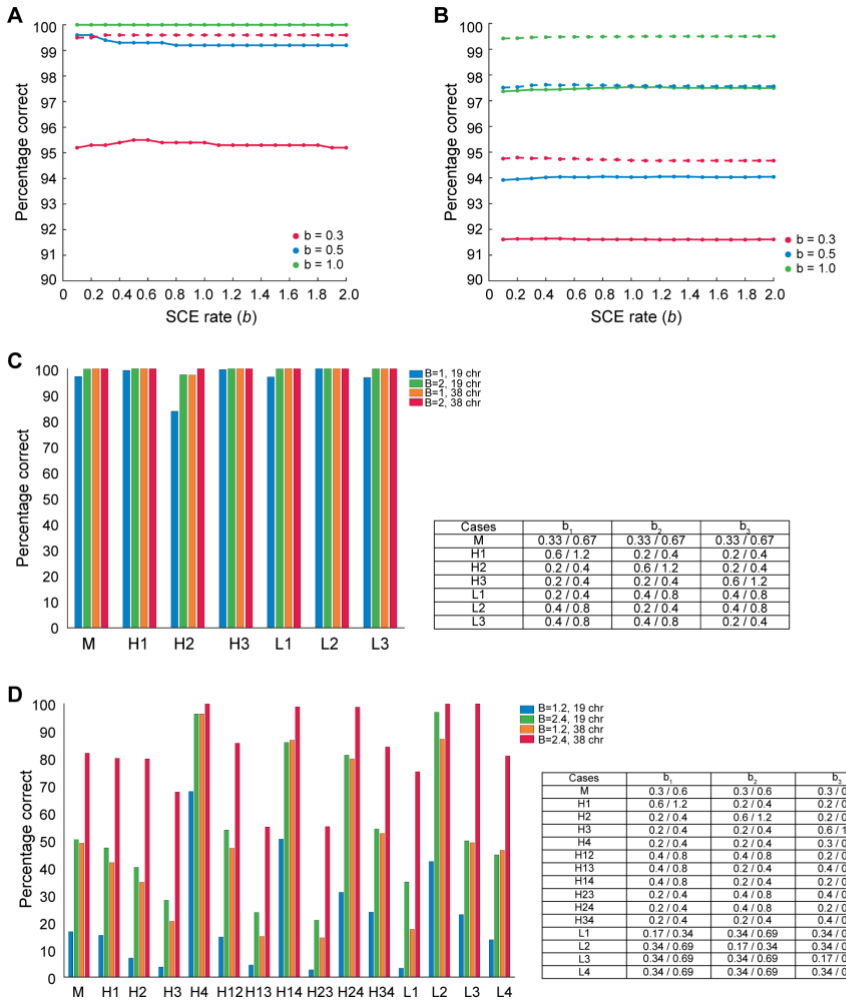


Figure S 1.4 Additional information increases the prediction accuracy of 32-cell trees at all subtree resolutions

(A) Panel shows the percentage of the full lineage, along with 2-, 4-, 8-, 16-cell subtrees, that are correctly predicted within simulated 32-cell trees as a function of SCE rates (*b*), assuming half of the sister pairs are known. This assumption is based on our recent work showing that strand-specific DNA methylation (5-methylcytosine or 5mC) can be used to identify half the sister cell pairs at the 32-cell stage of mouse embryogenesis (Sen et al., 2021 Nature Communications). The prediction accuracy is computed by simulating 2000 trees. Solid and dotted lines indicate cells where 5hmC can be quantified in 19 or 38 chromosomes, respectively.

(B&C) Panel shows the percentage of the full lineage, along with its subtrees, that are accurately predicted in simulated 32-cell trees as a function of SCE rates (*b*), where the rate of genomic variants is 0.3 (B) and 0.6 (C) per chromosome per cell division. The solid lines indicate the prediction accuracy using both 5hmC and gDNA information, while the dotted lines indicate the prediction accuracy using gDNA information alone. The prediction accuracy is computed by simulating 2000 19-chr trees.

(D) scH&G-seq using AluI, BseRI, or both enzymes enable detection of mitochondria DNA reads, while maintaining similar level of 5hmC detection as scAba-seq.



Cases	b_1	b_2	b_3
M	0.33 / 0.67	0.33 / 0.67	0.33 / 0.67
H1	0.6 / 1.2	0.2 / 0.4	0.2 / 0.4
H2	0.2 / 0.4	0.6 / 1.2	0.2 / 0.4
H3	0.2 / 0.4	0.2 / 0.4	0.6 / 1.2
L1	0.2 / 0.4	0.4 / 0.8	0.4 / 0.8
L2	0.4 / 0.8	0.2 / 0.4	0.4 / 0.8
L3	0.4 / 0.8	0.4 / 0.8	0.2 / 0.4

Cases	b_1	b_2	b_3	b_4
M	0.3 / 0.6	0.3 / 0.6	0.3 / 0.6	0.3 / 0.6
H1	0.6 / 1.2	0.2 / 0.4	0.2 / 0.4	0.2 / 0.4
H2	0.2 / 0.4	0.6 / 1.2	0.2 / 0.4	0.2 / 0.4
H3	0.2 / 0.4	0.2 / 0.4	0.6 / 1.2	0.2 / 0.4
H4	0.2 / 0.4	0.2 / 0.4	0.3 / 0.6	0.6 / 1.2
H12	0.4 / 0.8	0.4 / 0.8	0.2 / 0.4	0.2 / 0.4
H13	0.4 / 0.8	0.2 / 0.4	0.4 / 0.8	0.2 / 0.4
H14	0.4 / 0.8	0.2 / 0.4	0.2 / 0.4	0.4 / 0.8
H23	0.2 / 0.4	0.4 / 0.8	0.4 / 0.8	0.2 / 0.4
H24	0.2 / 0.4	0.4 / 0.8	0.2 / 0.4	0.4 / 0.8
H34	0.2 / 0.4	0.2 / 0.4	0.4 / 0.8	0.4 / 0.8
L1	0.17 / 0.34	0.34 / 0.69	0.34 / 0.69	0.34 / 0.69
L2	0.34 / 0.69	0.17 / 0.34	0.34 / 0.69	0.34 / 0.69
L3	0.34 / 0.69	0.34 / 0.69	0.17 / 0.34	0.34 / 0.69
L4	0.34 / 0.69	0.34 / 0.69	0.34 / 0.69	0.17 / 0.34

Figure S 1.5 scPECLR is robust to initial estimates of the SCE rate and to varying SCE rates at each cell division

(A&B) The sensitivity of the prediction accuracy to initial estimates of the SCE rate was tested for (A) 8-cell and (B) 16-cell trees. Trees were simulated with a constant SCE rate of $b = 0.3$ (red), $b = 0.5$ (blue), $b = 1.0$ (green). To test the robustness of the algorithm, instead of estimating the SCE rate from the data, values ranging from 0.1 to 2.0 were used during the first iteration of scPECLR to predict the tree. We found that the percentage of trees that were accurately predicted was robust across the range of SCE rates for cells containing both 19 (solid lines) and 38 chromosomes (dotted lines). Note that in panel (A), the dotted lines corresponding to $b = 0.5$ and $b = 1.0$ are hidden behind the solid line corresponding to $b = 1.0$, as they all have 100% prediction accuracy for all SCE rates. These results are based on 1000 simulated trees for each condition.

(C) Panel shows the percentage of 8-cell trees (with cells containing 19 or 38 chromosomes) that are accurately predicted for varying SCE rates over the 3 cell divisions. To systematically compare the prediction accuracy of scPECLR, the combined SCE rate ($B = b_1 + b_2 + b_3$) is held constant over all cell divisions. The results show that scPECLR can accurately predict the lineage of 8-cell trees even when the SCE rates vary with each cell division. The table denotes the SCE rates of each cell division for each condition.

(D) Panel shows the percentage of 16-cell trees (with cells containing 19 or 38 chromosomes) that are accurately predicted for varying SCE rates over the 4 cell divisions. M denotes cases where the SCE rate is constant over all cell divisions. H_i (and L_i) denotes cases where the SCE rate is higher (or lower) in the i^{th} cell division, and H_{ij} denotes cases where the SCE rate is higher in the i^{th} and j^{th} cell division than in the other two cell divisions. Again, the combined SCE rate is held constant. These results are based on 5000 simulated trees for each condition. The table denotes the SCE rates of each cell division for each condition.

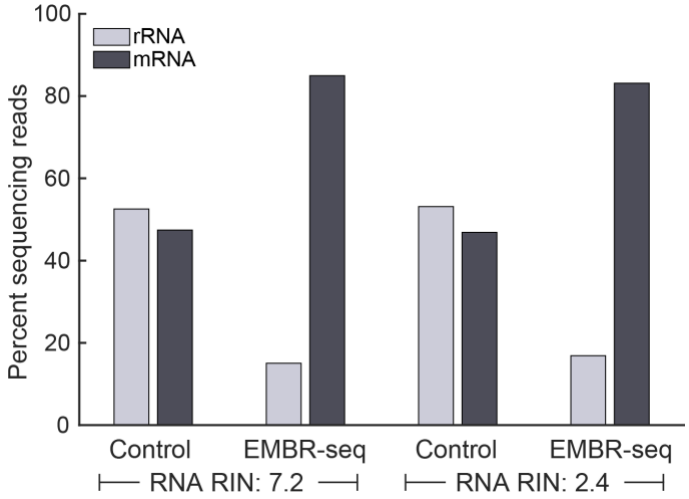


Figure S 2.1 EMBR-seq effectively depletes rRNA from fragmented total RNA.

EMBR-seq depletes rRNA to 15% and 17% of the mapped reads for total RNA samples with RIN scores of 7.2 and 2.4, respectively. In both cases, mRNA accounts for more than 80% of the mapped reads. These experiments were performed starting with 100 ng total RNA from *E. coli.*

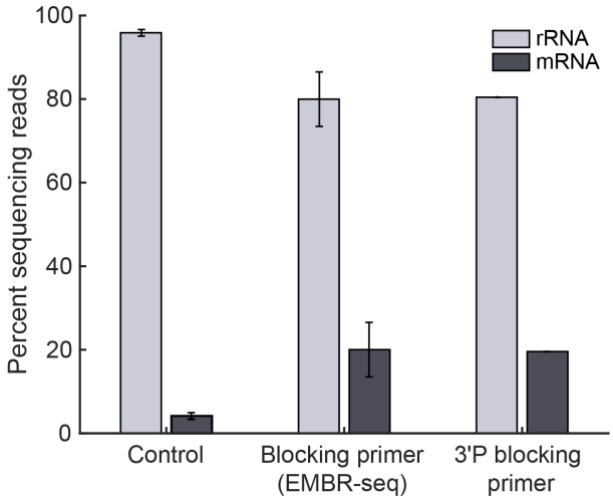


Figure S 2.2 Combining Terminator™ 5'-phosphate-dependent exonuclease (TEX) digestion with EMBR-seq does not improve rRNA depletion.

Performing TEX digestion prior to EMBR-seq results in less efficient rRNA depletion and mRNA enrichment compared to experiments without TEX (Figure 3.2A) ($n \geq 2$, except in TEX + 3'P blocking primer where $n = 1$). These experiments were performed starting with 100 ng total RNA from *E. coli.* Error bars represent standard deviations.

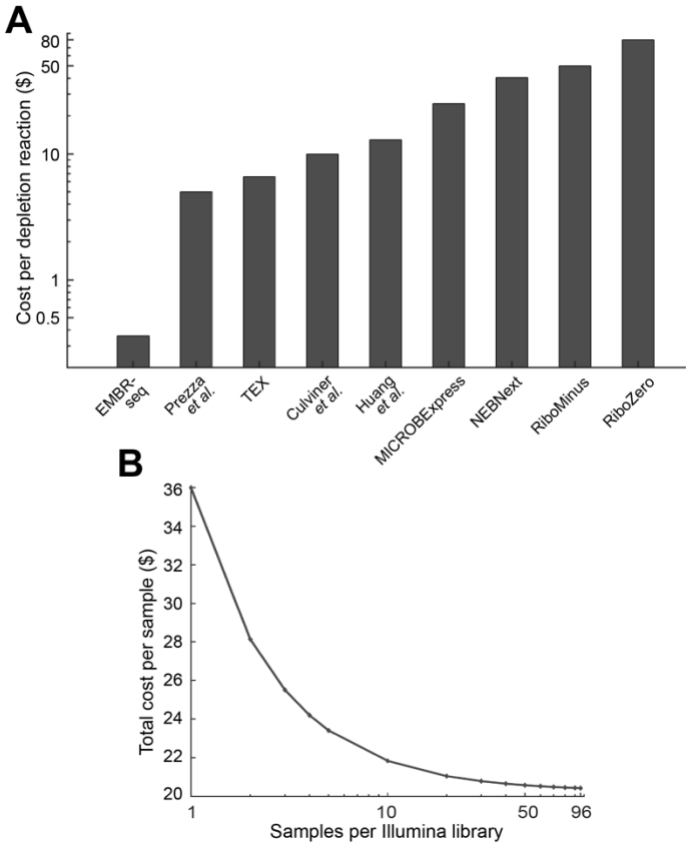


Figure S 2.3 Cost associated with performing EMBR-seq.

(A) The cost of performing rRNA depletion in EMBR-seq is ~$0.40 per reaction. The cost per reaction in EMBR-seq is an order of magnitude lower than other published rRNA depletion methods and commercial kits (Appendix E2: Tables S2.1, S2.2, S2.3). (B) The plot shows the total cost for the complete EMBR-seq protocol per sample (starting from total bacterial RNA extraction to Illumina library preparation) as a function of the number of samples multiplexed (using the sample barcodes in the RT primer) per Illumina library. Starting from 1 sample per Illumina library to 96 samples per Illumina library, the total cost drops from $36 to $20 per sample.

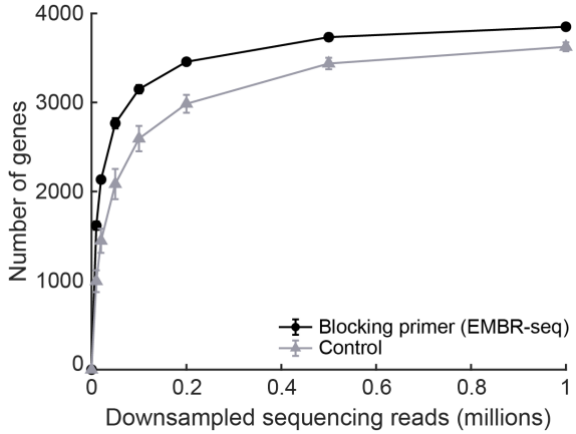


Figure S 2.4 Higher number of genes detected in EMBR-seq is not dependent on the sequencing depth.

To ensure that the number of genes detected in EMBR-seq samples compared to the control samples is not an artifact of sequencing depth, we downsampled the mapped sequencing reads to show that EMBR-seq

detects more genes at different levels of downsampling. The figure also shows that the number of genes detected does not increase substantially beyond ~0.5 million mapped reads, suggesting that our sequencing libraries have been sequenced at sufficient depth ($n = 3$). Error bars represent standard deviations. For the EMBR-seq group, error bars are of the same scale as the size of the data points.

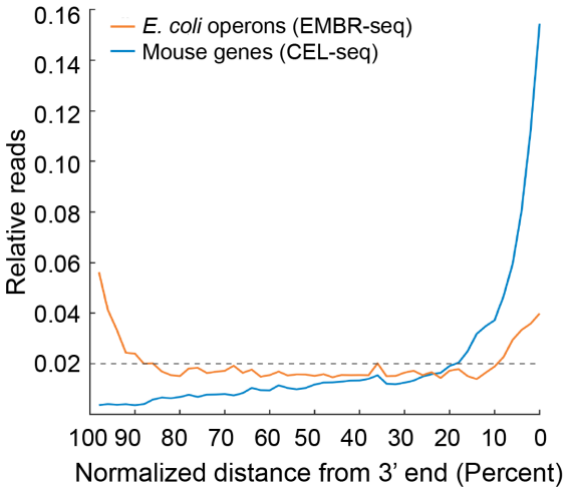


Figure S 2.5 Distribution of reads along *E. coli* operons in EMBR-seq.

The panel shows the distribution of reads along *E. coli* operons obtained from EMBR-seq and mouse genes obtained from CEL-seq. The normalized distance from the 3' end is based on discretizing the *E. coli* operons and mouse genes into 50 bins. The dotted line indicates the expected distribution of reads from each bin in the absence of any detection bias. The *E. coli* data is obtained from 100 ng starting total RNA.

Figure S 2.6 Gene transcript count correlation between different input total RNA amounts in EMBR-seq.

(A-C) Panels show gene transcript count Pearson correlations between 100 ng starting total RNA and lower input total RNA in EMBR-seq. As expected, the Pearson correlation drops when starting with lower amounts of total RNA. These experiments were performed with total RNA from *E. coli.*

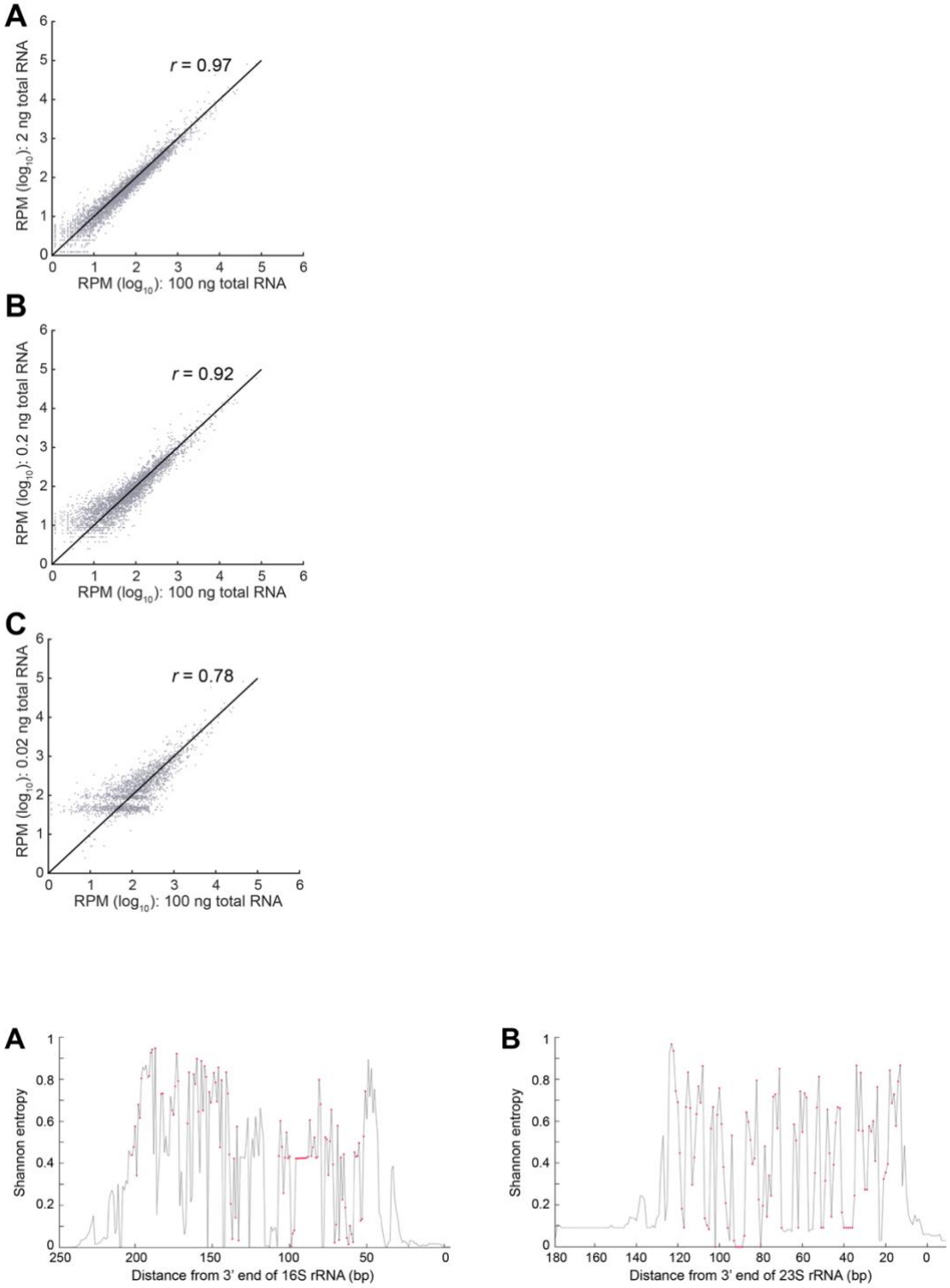


Figure S 2.7 Quantification of 16S and 23S rRNA sequence conservation using Shannon entropy.

(A,B) The panels show Shannon entropy scores for the sequence alignment of the last 100 bases of 16S and 23S rRNA from 4000 and 119 species, respectively. The red dots are the locations of the *E. coli* bases. The minimum entropy score of zero indicates that a position is completely conserved across all species analyzed, and the maximum Shannon entropy of 1.0 indicates that the bases at a location are uniformly distributed among species, or minimally conserved.

E. ***Supplementary Table***

1. Chapter 2

Location	Reference Base	SNP Base	% SNP	Total Sites	Found in HT-29?*	Found in TF-1?*
73	A	G	100.0	122	Y	Y
114	C	T	100.0	120	Y	
263	A	G	100.0	178	Y	Y
497	C	T	99.4	167	Y	
686	A	G	21.3	47		
710	T	C	20.0	50		
711	T	C	20.0	50		
750	A	G	100.0	129	Y	Y
1189	T	C	99.9	12942	Y	
1438	A	G	99.7	7343	Y	Y
1811	A	G	100.0	10	Y	
2706	A	G	100.0	218	Y	Y
3105	AC	A	100.0	6		
3480	A	G	99.5	222	Y	
4769	A	G	99.9	1926	Y	Y
7028	C	T	100.0	187	Y	Y
8860	A	G	100.0	91	Y	Y
9055	G	A	100.0	1143	Y	
9698	T	C	100.0	549	Y	
10398	A	G	100.0	53	Y	Y
10550	A	G	99.6	229	Y	
10978	A	G	100.0	723	Y	
11299	T	C	100.0	1150	Y	
11467	A	G	98.7	155	Y	
11470	A	G	100.0	154	Y	
11719	G	A	100.0	4679	Y	Y
11914	G	A	100.0	138	Y	
12308	A	G	100.0	92	Y	
12372	G	A	100.0	46	Y	
12954	T	C	99.8	15205	Y	
14167	C	T	99.6	2740	Y	
14766	C	T	99.6	4966	Y	Y
14798	T	C	99.9	8627	Y	

15326	A	G	99.9	61977	Y	Y
15924	A	G	100.0	86	Y	
16224	T	C	100.0	1272	Y	
16234	C	T	100.0	1192	Y	
16311	T	C	99.8	661	Y	
16519	T	C	100.0	17	Y	

*Diroma et al. 2020 Genes

Table S 1.1

SNPs detected in a H9 scH&G library, using BseRI and AbaSI enzymes

2. Chapter 3

Reaction Step	Reagent	Vendor	Catalog #	Cost ($)	Amount	Amount required	Unit	Cost ($/rxn sample)	Negligible costs
RNA extraction	Trizol	Thermo Fisher Scientific (Ambion)	15596018	326	200	0.25	mL	0.4075	Water; Ethanol
	Chloroform	UCSB chem store	1008-3862	11.66	500	0.05	mL	0.001166	
	GlycoBlue	ThermoFisher	AM9515	85	300	0.223333333	uL	0.063277778	
	Isopropanol	UCSB chem store	1022-2204	19.41	4000	0.125	mL	0.000606563	
Terminator digestion	Terminator (came with Buffer A)	Lucigen	TER51020	218	40	1	uL	5.45	Water
	RNAseOUT	ThermoFisher (Invitrogen)	10777019	171	125	0.5	uL	0.684	
	RNA beads	Beckman Coulter	A63987	760	40	0.025	mL	0.475	
	EDTA 100mM	Sigmal Aldrich	03690-100ML	38.31	500	0.001	mL	0.00007662	
PolyA addition	E coli polyA pol	NEB	M0276S	70	20	0.1	uL	0.35	Water; First strand buffer; ATP
	Blocking primers, 5S, no Pi	IDT	See *Primers* in Appendix B8	12	100	0.017	nmol	0.00204	
	Blocking primers, 16S, no Pi	IDT	See *Primers*	12	100	0.017	nmol	0.00204	
	Blocking primers, 23S, no Pi	IDT	See *Primers*	12	100	0.017	nmol	0.00204	
	Blocking primers, 5S, 5' Pi	IDT	See *Primers*	37	100	0.017	nmol	0.00629	
	Blocking primers, 16S, 5' Pi	IDT	See *Primers*	37	100	0.017	nmol	0.00629	
	Blocking primers, 23S, 5' Pi	IDT	See *Primers*	37	100	0.017	nmol	0.00629	
Pre-RT	RTprimer	IDT	See *Primers*	32.8	100	0.001	nmol	0.000328	

	Blocking primers, 5S, no Pi		IDT	See *Primers*	12	100	0.017	nmol	0.00204	
	Blocking primers, 16S, no Pi		IDT	See *Primers*	12	100	0.017	nmol	0.00204	
	Blocking primers, 23S, no Pi		IDT	See *Primers*	12	100	0.017	nmol	0.00204	
	Blocking primers, 5S, 5' Pi		IDT	See *Primers*	37	100	0.017	nmol	0.00629	
	Blocking primers, 16S, 5' Pi		IDT	See *Primers*	37	100	0.017	nmol	0.00629	
	Blocking primers, 23S, 5' Pi		IDT	See *Primers*	37	100	0.017	nmol	0.00629	
	dNTP 10mM		NEB	N0447L	236	4000	0.5	uL	0.0295	
RT	RNAseOUT		Thermo Fisher Scientific (Invitrogen)	10777019	171	125	0.5	uL	0.684	
	Superscript II		Invitrogen	18064-014	304	50	0.5	uL	3.04	
Second Strand Synthesis	Second stand buffer		Invitrogen	10812-014	153	500	12	uL	3.672	Water
	dNTP 10mM		NEB	N0447L	236	4000	1.2	uL	0.0708	
	E. coli ligase		Invitrogen	18052-019	49	10	0.4	uL	1.96	
	E.coli DNA polI		Invitrogen	18010-025	380	100	1.5	uL	5.7	
	E. coli RNaseH		Invitrogen	18021-071	461	60	0.4	uL	3.073333333	
	DNA bead clean up		Beckman Coulter	A63881	1200	60	0.06	mL	1.2	
IVT	MEGAscript T7 Transcription Kit (200 reactions)		Ambion	AMB13345	1168	400	1.6	uL	4.672	
Exo-SAP	Exo-sap It		Thermo Fisher Scientific (Affymetrix)	78200.200.UL	113	200	6	uL	3.39	
RNA fragmentation	Fragmentation buffer 200 mM		Made in House	N/A	2.38579437	10	0.0055	mL	0.001312187	Water
	Stop buffer 0.5 M EDTA		Made in House	N/A	38.31	100	0.00275	mL	0.001053525	Water
	RNA beads		Beckman Coulter	A63987	760	40	0.024	mL	0.456	
Library Preparation	pre-RT	RandomhexRT primer of cel-seq2	IDT	See Hashimshony et al. 2016	4.48	25	0.02	nmol	0.003584	
		dNTP 10mM	NEB	N0447L	236	4000	0.5	uL	0.0295	
	RT	RNAseOUT	Thermo Fisher Scientific (Invitrogen)	10777019	171	125	0.5	uL	0.684	

		Superscript II	Invitrogen	18064-014	304	50	0.5	uL	3.04	FS buffer and DTT
	PCR	NEBNext® High-Fidelity 2X PCR Master Mix	NEB	M05041L	360	6250	25	uL	1.44	Water
		RP1	IDT	Hashimshony et al. 2016	8	25	0.01	nmol	0.0032	
		Uniquely index RNA PCR primer	IDT	Hashimshony et al. 2016	25.2	100	0.01	nmol	0.00252	
	Bead clean up	Double-sided bead clean up	Beckman Coulter	A63881	1200	60	0.08	mL	1.6	
		Bead clean up	Beckman Coulter	A63881	1200	60	0.02	mL	0.4	
Bioanalyzer	Hgh Sensitivity DNA Kit		Agilent	5067-4626	580	110	1	sample	5.272727273	

Total Cost per depletion reaction (no Terminator, no Pi, includes BPs during RT step)	0.36224
Total Cost per depletion reaction (no Terminator, 5' Pi, includes BPs during RT step)	0.38774
Total Cost per depletion reaction (Terminator, no Pi, includes BPs during RT step)	6.97131662
Total Cost per depletion reaction (Terminator, 5' Pi, includes BPs during RT step)	6.99681662
Total Cost of one sample (excludes RNA extraction and bioanalyzer chip)	35.51537105
Total Cost of one sample (includes RNA extraction and excludes bioanalyzer chip)	35.98792139
Total Cost of one sample (includes RNA extraction and bioanalyzer chip)	41.26064866

Table S 2.1

Step-by-step EMBR-seq cost calculation

Pre-pool total cost	20.26
Post-pool total cost	15.72

Number of sample	Pre-pool	Post-pool	Total Cost per sample
1	20.26	15.72	35.99
2	20.26	7.86	28.13
3	20.26	5.24	25.51
4	20.26	3.93	24.20
5	20.26	3.14	23.41
10	20.26	1.57	21.84
20	20.26	0.79	21.05
30	20.26	0.52	20.79
40	20.26	0.39	20.66
50	20.26	0.31	20.58
60	20.26	0.26	20.53
70	20.26	0.22	20.49
80	20.26	0.20	20.46
90	20.26	0.17	20.44
96	20.26	0.16	20.43

Table S 2.2

Total EMBR-seq cost per multiple sample. The samples can be pooled at bead cleanup after second strand and pre-IVT step.

Supplier or publication	Kit or method name	Working principle	Cost/sample ($, approx.)	Minimal input	Percent rRNA left	Advantages	Disadvantages
EMBR-seq (Wangsanuwat et al., 2020)	EMBR-seq	Poly(A)-tail ing with rRNA blocking primers	0.36	20 pg	~10-22%	1. Inexpensive and small up-front cost 2. Simple design: 1 oligo per type of rRNA 3. Shown to work for starting material as low as 20 pg	1. Does not deplete as much rRNA as other methods 2. Minor detection bias
Prezza et al., 2020	DASH	Cas9-mediated cleavage of rRNA-derived cDNA	3-7	~0.4 ng	~10-50%	1. Shown to work for both gram-negative bacteria and anaerobic gut bacteria 2. Low cost 3. Has software for automated guide RNA design 4. Shown to work for starting material as low as 400 pg 5. Does not lower the amount of starting material for initial cDNA synthesis and PCR amplification	1. Requires more than 100 primer sequences, need around 650-800 oligos for more efficient depletion 2. Does not deplete as much rRNA as other methods 3. Depletion efficiency is variable and depends on the concentration ratio of sgRNA and Cas9 to the cDNA, which might require some optimization
Lucigen	Terminator exonuclease (TEX)	5' selective rRNA degradation	6.6	1 ug (recommended by Lucigen); 100 ng (this work); single bacteria cells (Kuchina et al., 2019)	~85-90% (He et al., 2010)	1. Simple enzymatic reaction that requires little design effort and is straightforward to perform in the lab	1. Poor rRNA depletion 2. Appears to worsen mRNA detection in combination with other techniques 3. Might introduce 5' detection bias

Supplier or publication	Kit or method name	Working principle	Cost/sample ($, approx.)	Minimal input	Percent rRNA left	Advantages	Disadvantages
Culviner et al., 2020	\	Oligo-based rRNA pull-down	10	2 ug	~20-25%	1. Has an algorithm for designing oligos for any species or combination of species of interest 2. Shown to work for at least 3 common bacterial strains	1. Requires multiple oligo optimization cycles *in silico*, up to 100 rounds 2. New optimization step might be required for every new combination of species 3. Up-front cost for the oligos 4. Only microgram-level starting material reported
Huang et al., 2019	\	RNase H-based	12.94	100 ng	<5% to ~25%	1. Shown to work for bacteria from 3 distinct phyla 2. Oligo probes can be applied to closely related species 3. Very good rRNA depletion 4. Has a simple tool to design probe libraries 5. Option to order probes or synthesize in house	1. Requires > 80 oligos 2. Up-front cost for the oligos 3. Lowest input reported is 100 ng
Thermo Fisher Scientific	MICROBExpress	Oligo-based rRNA pull-down	25	2 ug	1 to <10%	1. Excellent rRNA depletion 2. Works for many different gram-positive and gram-negative bacterial species	1. Cheaper alternate oligo-based pull-down methods available 2. Likely not easily customizable to other species not validated by the manufacturer 3. Only microgram-level starting material reported

Supplier or publication	Kit or method name	Working principle	Cost/sample ($, approx.)	Minimal input	Percent rRNA left	Advantages	Disadvantages
NEB	NEBNext rRNA depletion kit (bacteria)	RNase H-based	40.5	10 ng	<2%	1. Excellent rRNA depletion 2. Compatible with both gram-positive and gram-negative organisms, shown to work well across at least 20 different bacterial species	1. Cheaper alternate RNase H-based methods available 2. Likely not easily customizable to other species not validated by the manufacturer
Thermo Fisher Scientific	RiboMinus Transcriptome Isolation Kit, bacteria	Oligo-based rRNA pull-down	50	2 ug	<2%	1. Excellent rRNA depletion	1. Cheaper alternate oligo-based pull-down methods available 2. Likely not easily customizable to other species not validated by the manufacturer 3. Only microgram-level starting material reported
Illumina	RiboZero Plus rRNA depletion kit	RNase H-based	80 (reported by Prezza et al, 2020)	10 ng	<2%	1. Excellent rRNA depletion 2. Shown to work well across at least 25 different species 3. Works for both prokaryotic and eukaryotic samples	1.Cheaper alternate RNase H-based methods available 2. Likely not easily customizable to other species not validated by the manufacturer

Table S 2.3

Comparison of rRNA depletion methods.