STORM-seq reveals differentiation trajectories of primary human Fallopian tube epithelium

Benjamin K. Johnson[1]*, Mary Rhodes[1,2]*, Marc Wegener[2], Pamela Himadewi[1], Kelly Foy[1], Joshua L. Schipper[4], Rebecca A. Siwicki[2], Larissa L. Rossell[3], Emily J. Siegwald[3], Dave W. Chesla[3], Rachael T. C. Sheridan[4], Marie Adams[2], Hui Shen[1‡], Timothy J. Triche Jr.[1‡]

[1]Department of Epigenetics, Van Andel Institute, Grand Rapids, MI, USA
[2]Genomics Core Facility, Van Andel Institute, Grand Rapids, MI, USA
[3]Spectrum Health Accelerator of Research Excellence, Spectrum Health System, Grand Rapids, MI, USA.
[4]Flow Cytometry Core Facility, Van Andel Institute, Grand Rapids, MI, USA

*These authors contributed equally
‡These authors share senior authorship

Corresponding authors

Timothy J. Triche, Jr.
Van Andel Institute
333 Bostwick Ave. NE
Grand Rapids, MI 49503
tim.triche@vai.org

Hui Shen
Van Andel Institute
333 Bostwick Ave. NE
Grand Rapids, MI 49503
hui.shen@vai.org

Abstract

We present Single-cell TOtal RNA Miniaturized sequencing (STORM-seq), a full-length single-cell ribo-reduced RNA sequencing protocol, optimized to profile thousands of cells per run. Using off-the-shelf reagents and random hexamer priming, STORM-seq recovers comprehensive RNA profiles from single cells with library complexity approaching that of bulk RNA-seq. STORM-seq identifies thousands of additional coding and non-coding transcripts not detected by existing methods, and recovers clinically relevant structural variants at the single-cell level. We apply STORM-seq to primary human fallopian tube epithelium (FTE), a complex solid tissue key to both human reproductive biology and ovarian carcinogenesis. In differentiation trajectory analyses, the improved resolution from STORM-seq reveals intermediate/transitional cell states, and a putative progenitor cell population. The results support a trajectory from a bipotent progenitor population to ciliated and secretory cell types in normal FTE. These findings are consistent across human subjects, sequencing depths, and platforms. Long intergenic non-coding RNAs (lincRNAs) appear as key driver genes in both ciliated and secretory lineage trajectories, underscoring the importance of both coding and non-coding RNA in this tissue. By capturing essentially complete individual cellular transcriptomes, STORM-seq sheds new light on the transcriptional programs that establish cellular state and fate in complex tissues.

Introduction

Over the past decade, single-cell RNA sequencing (scRNA-seq) has yielded a revolution in dissecting the cellular heterogeneity of development and disease [1]. Via commercialization and improvement of existing protocols, the number of cells profiled and the number of features measured in each cell has risen continuously, such that each cell in a developing organism can now be profiled in parallel [2,3]. Early efforts captured polyadenylated transcripts from individual cells[4], using plates, microwells, or microfluidic emulsion with oligo-dT beads to segregate and label cells.

The most popular microfluidic techniques quantify genes by 5'/3'-end tags. These strategies scale profiling to millions of cells, controlling for amplification bias by incorporation of unique molecular identifiers (UMIs) [5]. However, profiling one end of the RNA fragments limits detection of alternative splicing or transcript isoform usage, and limited coverage of individual cells yields "dropouts," a phenomenon leading to unexpectedly high limits of detection for rare transcripts.

Full-length RNA-based protocols overcome such limitations by sequencing the entire RNA transcript, providing greater depth in each cell in exchange for fewer cells per unit of sequencing cost. These approaches were pioneered through the Switch Mechanism At the 5' end of RNA Templates sequencing (SMART-seq) protocol[6], and have culminated in the recently developed SMART-seq3[7], increasing gene- and isoform-level resolution while scaling to thousands of individual cells [8].

Recently, total RNA approaches have emerged to profile the total complement of RNA within individual cells after ribosomal RNA depletion. Some such approaches leverage full-length RNA protocols, relying upon enzymatic poly(A)-tailing of intact[9] or fragmented[10] total RNA prior to ribosomal depletion. However, the length of the poly(A)-tail can favor capture of transcripts with long tails[11], skewing and diminishing detection of non-coding RNA. Furthermore, decreasing concentrations of RNA and shorter RNA fragments promote long poly(A)-tails when using this strategy. This approach contrasts with traditional bulk ribo-reduced total RNA-seq, which uses random hexamer priming in place of oligo-dT-based capture and amplification. Furthermore, in existing total scRNA-seq techniques, custom fabrication of equipment and reagents are required to scale protocols[10]. Despite these limitations, both total scRNA-seq techniques serve as an important proof of concept: full-length whole-transcriptome profiling in single-cells is not only possible, but can reveal a complete picture of cell state. We developed STORM-seq to create a flexible, full-length, total scRNA sequencing protocol that leverages random hexamer priming, off-the-shelf reagents [12], and common lab automation equipment to scale to thousands of single cells. Here we show that this broadly applicable method can accelerate understanding of cellular heterogeneity and state in development and disease.

We also demonstrate STORM-seq's application to solid tissues with the human fallopian tube (FT). Also known as the uterine tube, oviduct, or salpinx, the FT is a highly specialized organ in the female reproductive tract essential for the transport of gametes, embryo and tubal fluid secretions. Aside

from its importance for reproductive biology, in the past two decades, it has been revealed that the FT likely hosts the cell of origin of the most common and deadly form of ovarian cancer, high grade serous ovarian carcinoma (HGSOC)[13–15]. The cellular and molecular makeup of the FT is thus of great importance for reproductive biology and gynecological malignancies alike. The innermost fallopian tube epithelium (FTE) is where precursors of HGSOC are thought to reside[15]. Well-established and characterized cell types in this epithelial layer include ciliated cells (FOXJ1+/CAPS+) and secretory cells (PAX8+/KRT7+) [13–15]. Additional cell types have been described[16], but less consistently. For example, intercalated 'peg' cells, characterized as EPCAM+/CD44+/ITGA6+, were proposed as stem-like or progenitor cells for the epithelial compartment [15]. However, these cells are poorly understood. Another cell type, dubbed 'basal'[17] or 'reserve' cells, is even less well characterized, recently proposed to likely represent resident immune cells [16–20].

Prior single-cell RNA sequencing studies of normal FTE have identified multiple populations of cells, including known types (ciliated and secretory), and a range of other cell types [16,18,19]. *Hu et al.*[18] used SMART-seq2 (poly-A-based) and identified multiple subpopulations within the secretory lineage, including a population of cells expressing an epithelial-mesenchymal transition (EMT) signature. Importantly, the presence of cells highly expressing this EMT signature in ovarian cancer patients was found to be associated with poor clinical outcomes[18]. More recently, *Dinh et al.*[5,19] used 10x Genomics to propose a de-differentiation trajectory of a population of $EpCAM^{Hi}$ (CD326), $CD44^{Hi}$, $THY1^{Hi}$ (CD90) secretory cells that give rise to ciliated cells through an intermediate cell population expressing *RUNX3*, challenging the notion of a bipotent progenitor for ciliated and secretory cells. This model, in turn, is challenged by more recent work[16]. Importantly, all three of the preceding studies rely on poly-A selection, with its limitations in both biotype and gene coverage. Here we apply Single-cell TOtal RNA Miniaturized sequencing (STORM-seq), a novel, full-length, ribo-reduced, single-cell total RNA sequencing technique, to reveal underlying dynamics of FTE differentiation.

Results

STORM-seq optimization and workflow

A schematic of the workflow for STORM-seq is presented in **Figure 1A**. STORM-seq is built upon the Takara SMART-seq Stranded kit, allowing use of off-the-shelf reagents and standard equipment. Briefly, single cells are index sorted into a 384-well microwell plate using fluorescence-activated cell sorting (FACS) containing a mixture of PBS and Takara Shearing Master Mix. After fragmentation, random primers allow for the capture of total RNA. Next, we leverage the SPT Mosquito HV liquid handler, which performs the addition of miniaturized volumes of First Strand Synthesis Master Mix and PCR1 Master Mix, generating cDNA and adding a unique sample/cell index sequence following PCR amplification. Though this protocol makes use of an automated liquid handler, all volumes can be hand-pipetted using a multi-channel pipettor. The SPT Mosquito HV then pools the 384 single cells into four pools of 96 cells. With a larger, pooled volume, the remainder of the protocol is then performed by hand. Each of the four pools are depleted of ribosomal RNA using Takara Bio's probe-based targeted enzymatic ribosomal depletion. The ribo-depleted 96-cell pools are amplified again and undergo a final paramagnetic bead cleanup. Cleaned, ribo-depleted pools are checked for quantity and quality using the Illumina iSeq and standard quality control metrics (e.g. Quantus and BioAnalyzer). Finally, the libraries are sequenced (**Figure 1A**).

As a proof-of-concept, initial optimizations were performed on the chronic myelogenous leukemia (CML) cell line K-562, given the breadth of available scRNA-seq on this cell line to compare with STORM-seq results. The first step in optimizing STORM-seq was ascertaining initial sort volume for single cells. Our approach, in contrast to the Takara SMART-seq Stranded Kit User Manual, began by combining PBS and Shearing Master Mix (hereafter referred to as sort Lysis Buffer). This optimization allowed for a sort volume of 2.2ul, increasing the throughput 4-6-fold (up to 576 cells per 96-reaction kit) compared to the original protocol and eliminating an unnecessary pipetting step. An additional benefit to sorting directly into Lysis Buffer is that the cells are immediately exposed to the RNase inhibitor contained in the master mix as they are sorted, stabilizing the RNA. This is important as premature lysis from sorting into PBS without RNase inhibitor could result in loss of template RNA. Next, we optimized the RNA fragmentation timing by decreasing the recommended incubation time from 6 minutes to 3 minutes (**Supplemental Figure 4**). Viable, single K-562 cells were sorted into one quadrant of separate 384-well plates containing Lysis Buffer and incubated for the indicated time points (**Supplemental Figure 2**). Prior experience on small numbers of cells suggested that longer than 6 minutes would lead to overly fragmented RNA in single cells (data not shown). Bioanalyzer traces from both 3 minutes and 6 minutes demonstrated that both time points yielded single-cell libraries. However, the 3 min time point showed more even coverage of fragment sizes across single cells, while the 6 minute is skewed toward smaller sizes (**Supplemental Figure 4B-C**). Visualization of raw single-cell sequence depth uniformity in Illumina sequencing analysis viewer following sequencing on the Illumina iSeq, we note that the 3-minute fragmentation gave a much more uniform coverage across single-cells relative to the 6-minute time point (**Supplemental Figure 4A**). Based on

both the library size difference and the evenness of the sequencing coverage, we adopted the 3-minute fragmentation time for STORM-seq (**Supplemental Figure 4A**).

Following fragmentation and cDNA synthesis, Illumina UDI sequences are added by the SPT Mosquito HV. It is ideal if minimal adapter dimer is present in the final library, or a large proportion of reads will be consumed sequencing these technical fragments. Our next optimization shows adding adapters at a third of the recommended concentration (Takara Bio SMARTer RNA UDI sets A and B) significantly reducing adapter contamination (see materials and methods; **Supplemental Figure 4A**).

One of the primary difficulties we encountered by decreasing reaction volumes and scaling to large numbers of cells in STORM-seq was performing ribosomal depletion due to timing of reagent additions and large numbers of single-cell pools. We modified the STORM-seq protocol to allow for larger pools of cells following fragmentation, cDNA synthesis, and single-cell indexing, compared to the original SMART-seq Stranded protocol. Per the SMART-seq Stranded User manual, cells are pooled into batches of 8-12 cells, meaning a 384-well plate would require up to 48-pools. With the miniaturized reaction volumes in STORM-seq, a 384-well plate can be pooled into batches of 96 cells, leaving 4 pools total to work with per plate. Given the volume limitations of the SPT Mosquito HV, the automated liquid handler generates the 4 pools by iteratively placing them into a column of a 96-well plate. For the remainder of the STORM-seq protocol, all steps are performed by hand. The large cell number pooling optimization is both theoretical and practical, decreasing possible technical effects of processing many smaller single-cell pools and lowering total hands-on time. An additional benefit to working with 4 pools of 96 cells per plate allows for larger volumes of beads to be used per pool during clean-up steps, enabling finer control over bead drying times to avoid over-drying beads as well as more complete removal of ethanol compared to on-plate cleanup strategies. We observe varying ethanol dry times, retention, and over-drying of beads when attempting to perform bead cleanups within a 384-well plate. We note that this optimized 4x96-cell pooling strategy yields high-quality libraries with no observable ribosomal peaks of major subunits as shown in the bioanalyzer traces from cryopreserved human fallopian tube epithelium (FTE) tissue (**Supplemental Figure 3**). Further, by pooling into 4 pools per plate, it enables processing of multiple 384-well plates simultaneously, scaling STORM-seq to thousands of single cells.

After finishing the STORM-seq library preparation, each of the pooled quadrants are checked for quantity and quality using a fluorescence-based readout (e.g. Quantus) and bioanalyzer traces. Additionally, pools can be sequenced using an Illumina iSeq, yielding additional information on library pool coverage prior to deep sequencing. Initial results suggested we had decreasing read depth per cell in the final columns of the 384-well plate, with up to a 33% dropout rate (**Supplemental Figure 5**). We overcame this issue by decreasing the pipetting speed of the automated liquid handler during viscous master mix additions and increasing the total dead volume in the reagent plates. The latter optimization was necessary due to retention of reagent (up to 100 nL) on the outside of the pipette tips, which were being changed in between every pipette motion to avoid contamination, leading to higher-than-expected consumption of reagents. Following these optimizations, we observe a cell dropout rate at or below 1% following initial quality control sequencing (**Supplemental Figure 5**).

STORM-seq detects thousands of coding and non-coding genes per cell

Given the significant optimizations and miniaturization of STORM-seq, built upon the SMART-seq Stranded kit, we sought to determine whether similar gene detection rates were retained compared to the parent kit, following the manufacturer protocol. We subsampled raw K-562 STORM-seq and SMART-seq Stranded fastqs to varying sequencing depths prior to alignment and tallied the number of genes detected per cell using a minimum threshold of greater than 0.1 transcript per million (TPM) (see materials and methods for details). Across all sequencing depths, we observe no loss of gene detection rates in STORM-seq K-562 cells compared to parent SMART-seq Stranded kit (**Figure 1B**). Further, we observe a modest increase in gene detection rates for STORM-seq at read depths greater than 0.5 million reads per cell, suggesting that the optimizations are beneficial. To understand how these gene detection rates compare to standard poly(A)-based SMART-seq2 and bulk total RNA, we subsampled raw fastqs to the same raw sequencing depths across technologies. On average, STORM-seq detects approximately 1,736 (SD = 168) more genes per cell compared to SMART-seq2 across sequencing depths at a TPM threshold of 0.1 in K-562 cells (**Figure 1B**). Compared to bulk total RNA-seq, STORM-seq detects approximately 4,700 (SD = 2,284) fewer genes across sequencing depths of 0.1-1 million reads, suggesting that STORM-seq provides complex libraries in single cells, approaching the complexity achieved in bulk methods (**Figure 1B**). Next, we compared STORM-seq, SMART-seq Stranded, and SMART-seq2 gene detection, stratified by annotated biotype. Across all gene biotypes, STORM-seq performs as well or better than SMART-seq Stranded and SMART-seq2 (**Figure 1C**). Further, STORM-seq identifies a similar set of protein-coding and lncRNA species compared to SMART-seq Stranded and SMART-seq2. STORM-seq collectively identifies 5951 more protein-coding and lncRNAs compared to the poly(A) SMART-seq2 (**Figure 1D**).

To compare STORM-seq to recently developed, bespoke single-cell total-RNAseq methods[9,10], we generated HEK293T single-cell libraries as a common cell type to compare across technologies. We compared STORM-seq and SMART-seq Total by sub-sampling each to 1 million raw reads per cell. Gene detection rates were compared across biotypes at a TPM detection threshold of 0.1. On average, STORM-seq detects 7,014 protein-coding genes (SD = 434) and SMART-seq Total detects 3,611 protein-coding genes (SD = 735). Additionally, STORM-seq detects 386 lncRNAs (SD = 67) and SMART-seq Total detects 257 lncRNAs (SD = 71). Notably, SMART-seq Total detects slightly more snRNA and miRNA species than STORM-seq (12 additional snRNA; 2 additional miRNA) (**Figure 1E**). STORM-seq represents the current state-of-the-art for single-cell ribo-reduced total RNA-seq.

STORM-seq detects clinically relevant fusions in individual cells

A particularly powerful aspect of RNA-seq is its ability to detect fusion transcripts. Some fusions (e.g. *BCR-ABL1* in chronic myeloid leukemia, or CML) are pathognomonic for disease. A subset of fusions are also targetable by small molecules (e.g. imatinib, which by targeting *BCR-ABL1* effectively slows the progression of CML). Thus RNA-seq can serve diagnostic as well as exploratory purposes.
A benefit to using K-562 cells for proof-of-concept is the presence of multiple, well-annotated fusion transcripts, including *BCR-ABL1*. This permits examining the sensitivity and specificity of RNA-seq

protocols for fusion transcript detection. We subsampled raw STORM-seq, bulk total RNA, and SMART-seq2 K-562 FASTQ files to 5 million reads, then used STAR-Fusion to detect fusion transcripts (see methods for details) [21]. Both STORM-seq and bulk total RNA detected *BCR-ABL1.* Approximately 84% (59/70) of STORM-seq K-562 cells had (on average) 4 junction spanning reads. Of note, this hallmark fusion was not detected in any SMART-seq2 libraries. One possibility is that poly-dT priming with low amounts of input RNA, as used in SMART-seq2 library preparation, diminishes detection of the BCR-ABL1 fusion. In contrast, both STORM-seq and bulk total RNA [22] utilize a random hexamer priming approach. SMART-seq2 detects two additional fusion transcripts, *CBX5-GTSF1* and *C16orf87-ORC6*. The *CBX5-GTSF1* fusion is not seen in previous K-562 characterizations, including the cancer cell line encyclopedia (CCLE)[23] and a recent comprehensive, phased reassembly of the K-562[24] genome with short and long reads. We conclude it is likely to be a false positive. In summary, STORM-seq can recover clinically relevant gene fusions in single cells.

STORM-seq of primary fallopian tube epithelium recovers known cell types and transitioning cells across patients and read depths

We applied STORM-seq to primary distal fallopian tube epithelium (FTE) from two patients and examined the ability of our technology to generate high-quality single-cell libraries from solid tissue, and define known cell types within the FTE across varying sequencing depths. Dissociated single cells were index sorted (DAPI dim, EpCAM+/CD238a-) into 376 wells of a 384-well plate with remaining wells filled with 4 positive and 4 negative controls (**Supplemental Figure 1** for gating scheme). Patient 1 was sequenced to a high depth (~5-20M reads/cell) and patient 2 sequenced at a lower depth (~500k-1M reads/cell) (**Figure 3A**). In total, we identify 308 and 333 high-quality cells for downstream analysis from patient 1 and patient 2. This demonstrates a robust FACS strategy and STORM-seq can generate high-quality single-cell libraries in FTE primary tissue (**Supplemental Figure 3**).

Index sorting is a powerful tool that complements single-cell transcriptome profiles from plate-based scRNA-seq protocols. Briefly, index sorting is the 1:1 association of single-cell flow cytometry fluorescence profiles (e.g. cell surface markers or functional dyes) to that of a given well or in a microwell plate. This alone makes plate-based scRNA-seq approaches, such as STORM-seq, an inherently multi-omic profiling technique through simultaneous measurements of cell features that can be directly associated with single-cell transcriptomes. As an example, to examine whether doublet or multiplet cells were present in our final sets of single cells, we plotted side scatter height, area, and width, colored by inferred clusters within each patient. Doublets or multiplets are expected to be outlier points shifted to the lower right quadrant in the side scatter area by height plots and shifted into the upper right quadrant of side scatter height by width plots. We observe a homogeneous distribution of singlets across clusters in both patients (**Supplemental Figure 7**). Further, we also verify that all cells retained for downstream analysis have cell surface expression and labeling of EpCAM across clusters and patients (**Supplemental Figure 8**). Taken together, the index sort information supports a successful gating strategy used to sort EpCAM+ single cells into each well of the microwell plate prior to STORM-seq library preparation (**Supplemental figure 1**).

To examine whether we capture known epithelial cell types found within normal FTE, we annotated within-patient cell cluster expression markers and compared them to known cell type expression markers from prior work [18,25]. Given the FACS gating strategy for EpCAM+ cells, most cells from both patients also express EpCAM at the RNA level (**Figure 3A**). Interestingly, this expression level varied by cluster. Additionally, secretory (*PAX8*, *KRT7*), ciliated (*FOXJ1*, *CCDC17*, *CCDC78*), and EpCAM+ infiltrating immune (*CD45*, *ITGAX*/CD11c, *CD14*) cells were also identified (**Figure 3A**), consistent with prior work [16,18,19,25]. However, we were not able to readily identify cell clusters corresponding to the less well-defined 'peg' cells using markers from a prior study[26] (*EpCAM*, *CD44*, *ITGA6*) within or across patients (**Supplemental Figures 9**). To maximize available information contained within this dataset, we integrated data from both patients using mutual nearest neighbors (MNN) (**Figure 3B**). All three cell types were retained (secretory, ciliated, and EpCAM+ immune cells) following integration based on marker gene expression (**Figure 3C**). Recent work from *Hu et al.*[18] also analyzed FTE using SMART-seq2 from four benign patient samples and characterized their respective cellular subtypes. To compare the identified secretory, ciliated, and basal/immune populations in this work, we reprocessed and integrated their SMART-seq2 data with the two patients characterized by STORM-seq with MNN. All expected cell types (secretory, ciliated, and basal/immune) are recovered in the joint dataset by marker gene expression, demonstrating the ability of STORM-seq to adequately identify expected cell populations of the distal FTE. Importantly, STORM-seq identifies more transitioning or intermediate cell types between secretory and ciliated populations compared to the four patients profiled using SMART-seq2 (**Figure 3D**).

Discovery of known and novel gene expression programs shaping FTE cell types

Joint marker gene analysis of both patients identified known and new gene expression programs distinguishing cell clusters, including a population of cells with epithelial-mesenchymal transition (EMT) signatures (**Figure 4A-B)**. Secretory cells represent the largest fraction of cells across both patients (~58%), mimicking the expected proportion *in vivo*. Within our data we observe the previously defined expression patterns distinguishing secretory cells from other cell types within the FTE (*EPCAM*, *PAX8*, *KRT7*, and *OVGP1*) (**Figure 4B**; **Supplemental table 1**) [16,18,19]. Top marker genes for the secretory clusters (clusters 5 and 7) include ovarian cancer-related genes (*MSLN*) and regulators of Wnt-signaling (*RSPO1*) (**Figure 4C**). Ciliated cell cluster marker genes (clusters 4 and 6) were enriched for cilia and flagella associated genes (*CFAP54*, *CFAP70*, *CFAP45*, *CFAP91*, *CFAP43*) and dynein genes (*DNAH9*, *DNAH2, DNAH6, DNAH10*) (**Figure 4D**). Consistent with the high ATP demand and mitochondrial association with beating cilia, we find mitochondrial respiratory complexes (*MT-ND1-5*) and ATP synthase subunits (*MT-ATP6*, *MT-ATP8*) as marker genes for ciliated cells [30–34]. Cluster 8 is immediately between cells transitioning to either ciliated or secretory cells and is referred to as the "intermediate" cluster. These cells possess low marker gene expression of both secretory and ciliated cells (**Figure 4B**). Consistent with this observation, marker gene expression is not exclusive to this population but tends towards either ciliated or secretory transitioning cells (clusters 1 and 2; **Figure 4B**; **Supplemental table 1**), suggesting an expression gradient corresponding to cell type differentiation. As an example, the ciliogenesis associated TTC17 interacting protein (*CATIP*) is identified as a marker gene for the "intermediate" cluster but is also expressed in the ciliated cell transitioning cluster 2 and terminally differentiated clusters 4 and 7

(**Figure 4E**). Conversely, the adenylyl cyclase 8 gene (*ADCY8*) plays a critical role in insulin secretion by modulating glucose homeostasis[35,36] and serves as a marker gene in the "intermediate" population, but is also expressed in the secretory lineage arm (**Figure 4E**). Cluster 9 expressed common lymphoid cell markers (CD45/PTPRC) and *RUNX3*, consistent with tissue resident T-cells and basal cells, as these cell also express EpCAM on their cell surface (**Figure 4F**; **Supplemental table 1**; **Supplemental Figure 8**). Cluster 3 expresses stem-cell genes (*GATA2*, *BMP2*), EMT signature genes (*TIMP3*, *ZEB1*, *DCN*, *SPARC*, *TWIST2*, *ACTA2*), and endothelial markers (*CAVIN2*, *PECAM1*, *VWF*) (**Figure 4G**; **Supplemental table 1**). Others[16] also identified a similar population of cells simultaneously expressing epithelial, EMT, and endothelial markers, and performed co-expression studies in tissue sections to identify whether these markers are found within the same cell or on different cells. Indeed, their results demonstrate that the epithelial:EMT and epithelial:endothelial markers can be co-expressed in the same cells [16]. Further, the EpCAM cell surface fluorescence index sort information from our data also confirm that these singlet cells express EpCAM on the cell surface, suggesting that they are not endothelial in origin (**Supplemental Figure 8**). As a putative progenitor population, given the expression gradients observed towards ciliated and secretory cells, we reasoned that cluster 3 may act as a common progenitor for both secretory and ciliated lineages.

Trajectory inference supports a bipotent progenitor model in FTE

To examine whether the observed expression gradient is part of a continuous differentiation trajectory to terminally differentiated cell types (e.g. secretory and ciliated), we applied RNA velocity[37,38], latent time[38], and pseudotime[39] analyses to patient-level data. RNA velocity infers a model of directed cellular dynamics given a static scRNA-seq gene expression profile through gene-level splicing kinetics, leveraging unspliced and spliced transcripts. Given that STORM-seq is a full-length random priming-based protocol (**Supplemental Figure 6**), we expect that the relative proportion of unspliced (nascent) transcripts detected will be higher than that of other protocols. Indeed, we observe a 59.5% ± 4.94% spliced (mature) transcript proportion and 40.5% ± 4.9% unspliced (nascent) transcript proportion in the FTE samples. This is higher than other protocols such as SMART-seq2/3 and 10x Chromium platforms, and comparable to VASA-seq [10]. We analyzed each patient independently using the stochastic RNA velocity model, with the immune populations excluded. Inferred cellular dynamics indeed support a continuous differentiation trajectory from cluster 3 in patient 1 and cluster 1 in patient 2 (hereafter referred to collectively as "root-like") populations of cells with distinct lineage commitments branching to ciliated and secretory cells (**Figure 5A; Supplemental Figure 11A**). To validate, we used an orthogonal method that does not rely upon spliced/unspliced transcripts called CytoTRACE[39] to estimate cellular plasticity and order cells from early/most plastic (e.g. stem-cell) to late/less plastic states (e.g. terminally differentiated cell type). Indeed, we observe a similar result to the RNA velocity inference with the earliest/most plastic cell types originating with the "root-like" populations of cells and defining distinct lineage commitments ending with terminally differentiated secretory and ciliated cell types (**Figure 5BC; Supplemental Figure 11BC**). Recent work[38] has suggested that so-called "latent time" derived from dynamical modeling of RNA velocity more accurately recapitulates real time compared to pseudotime[40]. Thus, as a third measure of trajectory reconstruction in FTE, we applied latent time to temporally reconstruct lineage commitment in patient 1. We observe a stepwise reconstruction of lineage fate consistent with both RNA velocity and

pseudotime (**Figure 5D**). Next, we identified key driver genes of lineage commitment towards either the secretory or ciliated branches using the previous RNA velocity inference. Top driver genes for ciliated (ULK4 and LINC01088) and secretory (PAX8 and LINC00937) were comprised of both coding and non-coding transcripts, underscoring the utility of profiling total RNA in this tissue (**Figure 5EF**). The relative expression and velocity of *CD44* are restricted to the progenitor-like cells and secretory lineage, indicating a possible role for *CD44* in secretory lineage fate (**Figure 5G**). Taken together, using multiple measures of trajectory inference, we observe a model supporting a bipotent progenitor with distinct lineage commitments towards ciliated or secretory cell types.

Discussion

STORM-seq is optimized and built upon the Takara SMART-seq Stranded kit. This enables users to perform the entire method using off-the-shelf reagents and equipment. While we have optimized the protocol to work with the SPT Mosquito HV, STORM-seq is amenable to other liquid handlers that can accurately pipette within the range of 0.5-5 uL. This pipetting range is also possible using a well-calibrated multichannel pipettor if a liquid handler is not accessible. Key optimizations within the protocol, such as pooling into four 96-cell pools and reagent delivery, enables STORM-seq to scale to thousands of cells with minimal loss of input cells. One of the current limitations of the protocol is the relatively low diversity of available Takara UDIs, allowing a maximum of 196 cells to be pooled within a sequencing lane. Ongoing work is exploring the incorporation of 384 UDIs to further increase throughput and decreasing sequencing costs by allowing pooling of an entire 384-well plate into a single sequencing lane. With the ability of STORM-seq to profile total RNA and reconstruct gene fusions not found in other poly(A)-based single-cell protocols (SMART-seq2), it is tempting to speculate that STORM-seq may also be capable of sensitively detecting other expression signatures such as transposable elements and eRNAs. Further, given the flexibility afforded to plate-based protocols through upstream FACS, STORM-seq complements existing droplet-based protocols to deeply profile targeted cell populations. It is often of interest to simultaneously measure as many cellular parameters as possible from a single-cell to extract maximal information about that cell in a destructive method like scRNA-seq. With FACS, it is possible to directly associate intracellular functional dyes (e.g. aldefluor for ALDH activity) and cell surface protein expression in a 1:1 fashion with single-cell total RNA transcriptomes through index sorting, making STORM-seq an inherently multi-omic protocol and increasing the number of parameters that can be measured in a single-cell.

Total RNA profiling methods at single cell level have been performed at somewhat smaller scales[41], with the exception of two recent publications. SMART-seq Total[9] and VASA-seq[10] both employ ligation of poly(A)-tails to total RNA prior to downstream analysis; the former relies upon DASH[42] for Cas9-mediated ribo-depletion, while the latter sustains substantially higher ribosomal content in pursuit of completeness. STORM-seq instead uses a more traditional random hexamer priming approach, similar to traditional bulk total RNA sequencing methods, followed by Takara Bio ZapR to enzymatically reduce ribosomal RNA. An advantage to using a random priming approach is that it avoids the long enzymatic tailing step of the protocol, and possible bias introduced during capture and amplification of transcripts or short transcript fragments with long poly(A)-tails, yielding diminished gene detection. Indeed, when comparing STORM-seq and SMART-seq total from HEK293T cells, we observed higher rates of detection in protein-coding genes and lncRNAs, with comparable detection of other gene biotypes at the same raw sequencing depth (**Figure 1E**). Of note, VASA-seq incorporates unique fragment indices (UFIs), which are similar to unique molecular identifiers (UMIs) in that they can be used to reduce PCR amplification bias. While STORM-seq is a non-UMI-based protocol, recent work[43] introduced a transformation for SMART-seq2 (also a non-UMI protocol) that can approximately account for PCR amplification bias and future work includes incorporation of UMIs into STORM-seq to accurately account for this bias.

Applying STORM-seq to primary distal fallopian tube epithelium from two patients, we identify known cell types and characterize differentiation trajectories towards terminally differentiated secretory and ciliated cells from a “root-like” progenitor population. Recent work[16] profiled four benign FTE samples using SMART-seq2 and identified a small sub-population of cells expressing an EMT signature. Similarly, the “root-like” population we identify in our study also bears a similar gene expression signature (*TIMP3, SPARC, DCN, ACTA2*) providing additional evidence that this population of cells can exist within the normal FTE. However, we also observe additional stem-cell like gene expression (*GATA2* and *BMP2*) as well as endothelial markers (*CAVIN2*, *VWF*, *PECAM1*) in our population of cells. Recent work[16] demonstrates co-expression of both EMT markers with epithelial markers, and also endothelial markers with epithelial markers. Further, the index sort information provides evidence that the “root-like” progenitor population is unlikely to be endothelial in nature (**Supplemental Figure 8**). Instead, these cells may be capable of endothelial differentiation as well as giving rise to epithelial cell types. *BMP2* is part of the transforming growth factor beta (TGF-β) pathway and a potent differentiation factor[44]. Additionally, *BMP2* is known to play a role in stimulating proliferation of ovarian cancer stem cells (OCSCs)[45]. *GATA2* is a critical regulator of human hematopoietic stem cells[46–49]. Thus, expression of differentiation factors like *BMP2* and known stem cell regulators like *GATA2* in a “root-like” progenitor population supports the proposal that these cells act as a bipotent progenitor within the FTE.

Prior work has shown that *ULK4* is essential for ciliogenesis, and *LINC01088* has been implicated as a tumor suppressor in ovarian cancer, with reduced expression correlating with poor patient outcomes [50–52]. This is consistent with a model of early loss of ciliated cells and a concomitant increase in secretory cells during disease progression of HGSOC and may be one possible mechanism shifting lineage commitments towards the secretory branch. The PAX family of nuclear transcription factors are known to drive lineage commitment in early embryogenesis with mouse studies demonstrating *PAX8* being tissue restricted to several developing tissues, including the Müllerian tract [25,53]. Here we show evidence *PAX8* expression may function as a driver of secretory lineage commitment in human FTE (**Figure 5F**). Another driver gene of secretory lineage commitment, *LINC00937*, has been associated with cervical cancer progression, though specific molecular functions remain poorly understood. Recent work from *Dinh et al.*[19] proposed a model where early secretory-like cells with an EMT-like signature undergo dedifferentiation through an intermediate population in a *RUNX3*-mediated gene program. In our data, we do not observe *RUNX3* functioning as a driver gene of positive velocity from secretory to ciliated cell states (**Figure 5G**). Prior studies[26] have shown *CD44* as a positive marker for peg cells in combination with *ITGA6* as a putative stem-cell within the FTE. *Dinh et al.* proposes a model of de-differentiation through an intermediate population of cells expressing *RUNX3*. In our work, we do not observe a dedifferentiation trajectory, and *RUNX3* expression is predominantly expressed in our “immune” cluster of cells, in line with Hu[18]. The ‘early secretory’ population of cells identified in *Dinh et al.* also express similar markers as the EMT population in *Hu et al.* and our “root-like” population, suggesting that common cell types are emerging from single-cell analyses of FTE. One explanation for these discrepancies emerges from Dinh et al., where pseudotime alone was used to reconstruct trajectories from 5’ poly-A transcriptomes averaging

700-1000 genes per cell. Recent work[54] has shown that both velocity and pseudotime based methods can yield novel insights depending on the experimental design, cellular coverage, and time scales. Additionally, several lower-abundance snoRNAs are identified as marker genes within the secretory clusters (*SNORA73A*, *SNORA73B*, *SNORA81*, *SNORA62*, *SNORA24*, *SNORD1C*) (**Figure 4B**; **Supplemental table 1**). Recent work demonstrated that SNORA73A and SNORA73B are critical for metabolic stress responses and regulate cell metabolism through mTOR [27,28]. Given the central role mTOR plays in nutrient sensing to control cellular metabolism in conjunction with secretory cells producing nutrient rich fluid for gametes[29], the SNORA73AB/mTOR axis may provide a pathway for maintaining secretory cell metabolic homeostasis and nutrient secretion. Especially when investigating rare cell populations, we propose that the additional resolution gained from deep total RNA profiling in single cells is a valuable tool to reconcile divergent observations.

In this work, we used cryopreserved (rather than fresh) cells. Nevertheless, recent work[16] using the 10x Chromium platform to profile fresh human FTE, and integration with SMART-seq2 data from *Hu et al.*[18], corroborate the majority of our findings. Additional experimental work is needed to validate the progenitor identity of the “root-like” population of cells, and its contributions (if any) to development of pro-oncogenic states linked to ovarian cancer.

Taken together, STORM-seq represents a flexible and broadly accessible method to profile ribo-reduced total RNA in single cells. Using this method, we are able to reconstruct the expected tissue heterogeneity and identify a “root-like” progenitor population of cells in primary human fallopian tube epithelium. Additionally, we infer cellular differentiation trajectories and propose this “root-like” population is likely a bipotent progenitor, capable of giving rise to secretory and ciliated epithelial cell types using both RNA velocity and RNA velocity-independent methods.

Materials and Methods

Cell lines and culture conditions

The K562 cell line was purchased from the American Type Culture Collection (ATCC; CCL-243). Cells were cultured in RPMI 1640 medium supplemented with 10% FBS and 1% Penicillin/Streptomycin (Gibco). HEK293T cells were purchased from the American Type Culture Collection (ATCC; CRL-3216) and cultured in DMEM medium supplemented with 10% FBS, 2 mM L-glutamine, and 1% Penicillin/Streptomycin (Gibco). Cells were maintained at 37°C in 5% CO_2 and sub-cultured at 1:5 ratio every 3-4 days.

Human Fallopian tube epithelium sample acquisition

Fallopian tube fimbriae were collected at Spectrum Health Hospital in Grand Rapids, Michigan. After initial screening for benign conditions not affecting the epithelium of the fallopian tube (e.g. uterine fibroids), patients undergoing total hysterectomy with salpingo-oophorectomy were provided written informed consent for sample collection and analysis. The study was approved by Institutional Review Board (IRB) of the Van Andel Research Institute under IRB #19017. Following surgical resection, the tissue was placed in saline and the single-cell dissociation protocol was started within 20 minutes of receiving the tissue.

Fallopian tube epithelium dissociation to single cells

Dissociation of fresh Fallopian tube epithelium was done using the Milltenyi gentleMACS Octo Dissociator with heaters and the human Tumor Dissociation Kit (Miltenyi Biotec). Briefly, the tissue was removed from the DMEM/RPMI 1640 media and manually minced in a sterile Petri dish. Minced tissue was placed into Milltenyi gentleMACs C tubes, complete with 4.7 mL of DMEM (Gibco) and supplemented with enzymes A, H, and R at the manufacturer recommended working concentrations (Miltenyi Biotec). Next, the tissue was dissociated into a single-cell suspension using the “soft tumor type” program (37C_h_TDK_1) on the gentleMACS Octo Dissociator. Following conclusion of the dissociation program, sample tubes were centrifuged at 300 x g for 3-5 minutes to pellet the cells. The supernatant was removed and the pellet resuspended in filter-sterilized flow sorting buffer (HBSS + 2% FBS + 25 mM HEPES). Remaining intact tissue was removed by passing the suspension through a Milltenyi SmartStrainer (Miltenyi Biotec) into a new 50 mL Falcon tube (Corning). Viable cells were counted using Trypan blue and the BioRad TC20 Automated Cell Counter. The remaining, strained single-cell suspension was transferred to 2 separate 15 mL Falcon tubes (Corning) and centrifuged to pellet cells (300 x g for 3-5 minutes). Pellets were resuspended in freezing media (90% FBS + 10% DMSO) and cryopreserved.

K562 and HEK293T cell sorting

Cells were spun down at 300g for 5 minutes and washed in DPBS (Gibco). K562 cells were resuspended in 100 uL of DPBS containing Zombie aqua fixable viability dye (Biolegend) at 1:1000 ratio, per 1 million cells. K562 cells were stained at room temperature for 15 minutes, quenched with 3 volumes of DPBS containing 2% FBS for 3 minutes, washed, and resuspended in 300 uL of DPBS

containing 2% FBS per 1 million of cells. HEK293T cells were resuspended in 300 uL of flow buffer (HBSS [Gibco], 2% FBS, 25 mM HEPES) containing 2 µg/mL DAPI. One, single, live, cell was index sorted into each well of either a 96-well plate containing 7 uL of DPBS (for "full volume" SMART-seq stranded libraries) or a 384-well plate containing 1x Takara lysis buffer (STORM-seq) using a MoFlo Astrios cell sorter running Summit v6.3 (Beckman Coulter). The Astrios was equipped with a 100 um nozzle at 25 psi. Abort mode was set to single and drop envelope to 0.5. A full gating strategy can be found in **Supplemental Figure 2**.

Fallopian tube epithelium cell sorting

Cryopreserved primary cells were rapidly thawed in a 37 °C water bath. The thawed cells were gently resuspended in several volumes of flow sorting buffer (HBSS + 2% FBS + 25 mM HEPES) and pelleted at 300 x g. Cell pellets were resuspended in 1 mL of flow sorting buffer and viable cells were counted using trypan blue and the BioRad TC20 Automated Cell Counter (BioRad). Cells were pelleted and resuspended in 100 uL of flow sorting buffer and stained for 20 minutes in the dark with 1.25 uL of EpCAM-PE (CD326; clone 1B7; Invitrogen) and CD235a-FITC (clone HIR2 (GA-R2); Invitrogen). Cells were pelleted and washed twice with 1 mL flow sorting buffer, with the final wash resuspending cells in 300 uL of sorting buffer. To assess viability, DAPI (Thermo Fisher) was added to a final concentration of 2 µg/mL to stained cells without washing. Single, live, EpCAM+, CD235a- cells were index sorted into designated wells of a 384-well plate containing 1X Takara SMART-seq stranded Shearing Master Mix. The Astrios was equipped with a 100 um nozzle at 25 psi. Abort mode was set to single and drop envelope to 0.5. A full gating strategy can be found in **Supplemental Figure 1**.

Bulk total RNA isolation and library construction from K562 cells

Total RNA from 5 million K562 cells were isolated using TRIzol (Invitrogen). Libraries were prepared by the Van Andel Genomics Core from 500 ng of total RNA using the KAPA RNA HyperPrep Kit with RiboseErase (v1.16) (Kapa Biosystems, Wilmington, MA USA). RNA was sheared to 300-400 bp. Prior to PCR amplification, cDNA fragments were ligated to Illumina UDIs. Quality and quantity of the finished libraries were assessed using a combination of Agilent DNA High Sensitivity chip (Agilent Technologies, Inc.) and QuantiFluor® dsDNA System (Promega Corp., Madison, WI, USA).

Single-cell K562 library construction

The STORM-seq method is adapted from the SMART-seq Stranded (SSS) kit (Takara Biosystems USA), currently the only commercially available total RNA preparation for single cells. STORM-seq modifies the SSS random priming protocol for miniaturization, improved efficiency, and throughput. Single-cell K562 libraries were prepared using the SSS Kit, following the manufacturer instructions with modifications. Following sorting into PBS, Takara Shearing Master Mix was added to all samples to a final concentration recommended by the kit. Half of the samples underwent 6 minute fragmentation time at 85 °C while the other half underwent 3 minute fragmentation time at 85 °C. Following first strand synthesis, the two sets of samples were processed together in a single batch to minimize technical effects. Section B step 4 in the protocol was modified to account for the incorporation of UDIs (Takara SMARTer RNA Unique Dual Index Kit-96A). A total of 2 uL of each UDI

was used instead of the recommended 1 uL each of the 5' and 3' PCR primers. For PCR1, 10 cycles of amplification were performed, followed by pooling according to Appendix A. The remainder of the protocol was performed according to manufacturer recommendations.

Miniaturized K562 and Fallopian tube epithelium library construction

Single cells were prepared using the SMART-Seq Stranded Kit (Takara USA) on the SPT Labtech Mosquito (HV) at 1:6, then 1:4 dilutions relative to the recommended volumes for subsequent steps in the protocol. In contrast to the manually prepared K562 libraries, for the miniaturized protocol, cells were sorted into a 384-well plate (Eppendorf twin-tec) containing a combination of Shearing Master Mix plus PBS and immediately transferred to a thermocycler and fragmented for 3 minutes at 85 °C. Following 1st Strand Synthesis, PCR1 (10 cycles of amplification) was performed with the addition of SMARTer RNA Unique Dual Index Set A (Takara SMARTer RNA Unique Dual Index Kit-96A) instead of Takara 5' and 3' PCR Primers. Samples were pooled across rows into groups of 12 cells according to Appendix A: Pooling Strategy for Single-Cell Applications and depletion of ribosomal cDNA was performed as indicated. 13 cycles of amplification were performed for PCR2 and the final library was eluted in 13ul 10mM Tris-HCl pH 8.0.

Sequencing

The manually prepared K-562 single cells were pooled and sequenced using 2x75 bp sequencing on an Illumina NextSeq 500 (Illumina Inc., San Diego, CA, USA). Base calling was done by Illumina NextSeq Control Software (NCS) v2.0 and output of NCS was demultiplexed and converted to FASTQ format with Illumina Bcl2fastq v1.9.0. Bulk and STORM-seq K562 libraries were pooled and 2x50 bp sequencing was performed on an Illumina NovaSeq 6000 sequencer using a 100 bp S2 sequencing kit (Illumina Inc., San Diego, CA, USA). Patient 1 fallopian tube epithelium STORM-seq libraries were also sequenced on a NovaSeq 6000 sequencer using a 2x100 bp S2 sequencing kit. Patient 2 fallopian tube epithelium STORM-seq libraries were sequenced on a Hiseq 4000 using a 2x150 bp sequencing kit at Fulgent Genetics. Base calling was done by Illumina NovaSeq Control Software (NCS) v2.0 and output of NCS was demultiplexed and converted to FASTQ format with Illumina Bcl2fastq v1.9.0. Base calling for the HiSeq sequencing run was performed at Fulgent Genetics and returned as FASTQ files.

Additional data sets used

The following publicly available datasets were downloaded and used for comparison to STORM-seq: 1) K562 SMART-seq2 (Illumina runs) - SRP132313, 2) HEK293T SMART-seq total - GSE151334 , 3) Normal FTE (benign donors) SMART-seq2 - SRP226182.

Dataset comparison

Both publicly available SMART-seq2 and internally generated STORM-seq K562 datasets were sequenced to a depth of approximately 5-7 million reads per cell. Publicly available HEK293T data were sequenced to a depth >1 million reads per cell on average. For platform comparisons, each cell was subsampled to 100,000, 500,000, and 1 million raw reads with seqtk, setting a seed of 100. Bulk total RNA and STORM-seq K562 cells had a subset of cells that were sequenced to >20 million reads

and were subsampled to 20 million reads using seqtk with a seed of 100. Subsampled reads were realigned to hg38 using salmon with modifications as described below.

Trimming STORM-seq and SMART-seq stranded libraries

STORM-seq library generation is based upon the Takara Bio SMART-seq stranded kit, which requires trimming of 3 bp from the 5' end of read 2. Prior to alignment and expression quantification, STORM-seq and SMART-seq stranded libraries were trimmed using TrimGalore (v0.6.3) using default parameters with the following modifications: 1) --trim-n, 2) --length 36, 3) --paired, 4) --clip_R2 3, and 5) --fastqc.

Trimming SMART-seq total and SMART-seq2 libraries

SMART-seq total HEK293T cells were trimmed using TrimGalore (v0.6.6) following a similar strategy as previously described [CITE]. Briefly, single-end reads were trimmed using default parameters with the following modifications: 1) --trim-n, 2) --length 18, 3) specifying two adapter sequences as -a AAAAAAAAAA -a TTTTTTTTTT, and 4) --fastqc. SMART-seq2 K562 data were trimmed using TrimGalore (v0.6.3). Default parameters were used with the following modifications: 1) --trim-n, 2) --length 36, 3) --paired, and 4) --fastqc.

Mapping and gene expression quantification of STORM-seq and SMART-seq stranded libraries

K562, HEK293T, and cryopreserved Fallopian tube epithelium (FTE) trimmed FASTQ files were aligned to GRCh38 (soft-masked primary assembly releases 101 for cell lines and 102 for FTE) with salmon (v1.3.0) using default parameters with the following modifications: 1) library orientation to ISR, 2) turning on the seqBias, gcBias, and posBias options due to the random priming approach in STORM, and 3) turning on the softclip option in selective-alignment mode. Gene expression was quantified using salmon and Ensembl gene annotations corresponding to the respective assembly release. Transcript counts were imported using the import_plate_txis() function in velocessor (v0.15.27 - https://github.com/trichelab/velocessor) in R (v4.0), constructing a SingleCellExperiment object. For RNA velocity inference, the FTE samples were mapped to a spliced/unspliced index generated by eisaR (v1.2.0) and salmon (v1.3.0) using Ensembl 102 assembly and annotations. Briefly, due to the full-length protocol of STORM-seq, we set featureType to extract "spliced" and "unspliced" transcript ranges at the getFeatureRanges step. All other steps were followed as described in the eisaR vignette for generating spliced/unspliced transcript files for indexing by salmon. After importing, transcript-level counts were collapsed to gene level. Transcripts per million (TPM) computed by salmon were also transformed using the izar_transform() function in velocessor as previously described[43] for non-UMI-based protocols like SMART-seq2 and STORM-seq. The SingleCellExperiment object containing the spliced and unspliced gene-level counts were exported as an h5ad file using zellkonvertor (v1.0.3) for import using scvelo[38]. Quality control and normalization were performed using the scran (v1.16.0) and scater (v1.16.0) packages in R. Briefly, low quality cells were removed using the perCellQCMetrics() and quickPerCellQC() functions in scater with default parameters. Library size normalization was done using the pooling and deconvolution approach implemented in scater[55]

Gene fusion calling in K562 bulk and single cells

K562 STORM-seq, SMART-seq2, and bulk total RNA were aligned using the STAR-Fusion pipeline (v1.8.1) and pre-built STAR-Fusion CTAT reference genome GRCh38_gencode_v32_CTAT_lib_Dec062019.plug-n-play. Specifically, for SMART-seq-style sequencing runs, recommended best practices were followed according to https://github.com/STAR-Fusion/STAR-Fusion/wiki/STAR-Fusion-scRNA-seq and https://github.com/STAR-Fusion/STAR-Fusion/issues/90. Following alignment and joint fusion calling, single-cell fusion calls were deconvolved using the starF_partition_final_by_sc.pl script provided with STAR-Fusion. Each single cell was then analyzed for fusions. Results were cross-referenced with deep phased *de novo* assemblies from Zhou et al[24]. Structural variants not appearing in CCLE or Zhou et al. were considered false positives for the purpose of comparisons.

Single-cell clustering in primary fallopian tube epithelium

Following library size normalization, the top 10% highly variable genes across filtered cells were retained and embedded using the runPCA() function in scater, setting *ncomponents* to 20. The number of principal components to retain were estimated using the scater function getClusteredPCs(). In total 7 PCs were retained for patient 1 (high-depth sequencing) and 15 PCs were retained for patient 2 and downstream clustering. Cluster analysis was performed using the retained PCs and building a shared nearest neighbor graph with the buildSNNGraph() function in scater. Clusters were identified using the cluster_walktrap() function in igraph (v1.2.6). Both t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold and Projection (UMAP) 2-D embedding and projections were used for visualization of the data. To compute t-SNE embedding and projections, the runTSNE() function in scater was used with 20 PCs for both patient samples. For computing density preserving UMAP (densMAP) embedding and projections, the densvis package (v1.00.6) and the densmap() function, setting *n_neighbors* to 15, *n_components* to 3, and metric to “euclidean”, was used with 20 PCs for both patients. Clusters inferred as described above were then colored on t-SNE and densMAP projections.

Integration of multi-patient primary fallopian tube epithelium STORM-seq

Integration of both patient 1 and patient 2 STORM-seq data, as well as integration with SMART-seq2 FTE data, was performed using mutual nearest neighbors as implemented in batchelor (v1.6.3) using the fastMNN() function as described in Orchestrating Single-Cell Analysis with Bioconductor[56]. Briefly, for STORM-seq integration, multi-batch normalization was performed using the multiBatchNorm() function in batchelor, scaling towards the lowest coverage batch - in this case is patient 2. Next, feature selection was performed by averaging the multi-patient variance components using the combineVar() function in scran (v1.16.0) and keeping genes above the trend as previously described[56]. In total, 17,573 genes were retained. Next, the rescaled patient counts and selected features were passed to fastMNN() using default parameters with the following modifications: 1) k = 15 and 2) subset.row = selected features described above. The integrated data were then used to compute clusters and a 2-D t-SNE projection similar to the process described above, with the only difference being that 50 PCs were used as input to the runTSNE() function. To integrate STORM-seq data with previously published SMART-seq2 FTE data, the process described above was repeated

for integration, treating batches as 1) patient 1 STORM-seq, 2) patient 2 STORM-seq, 3) SMART-seq2 FTE data.

Cell type annotation in primary fallopian tube epithelium

Cell types were defined using the clusters inferred as described above, and both known cell type markers and inferred marker genes.

RNA velocity of primary Fallopian tube epithelium

Single-cell experiment objects were exported as an h5ad file using zellkonverter (v1.0.3). Exported h5ad files were imported into python (v3.8.5) using scanpy (v1.7.1) and scvelo (v0.2.3), converting to an AnnData object. The clusters containing immune cells were excluded for velocity inference. Each patient was processed separately, but using the same parameters for initial preprocessing and model fitting. Briefly, spliced and unspliced counts were filtered and normalized using the filter_and_normalize() in scvelo with the following modifications: 1) min_shared_counts=20 and 2) n_top_genes=1000. Next, first- and second-order moments were computed using the moments() function with the following modifications: 1) n_pcs=5 and 2) n_neighbors=15. The number of PCs and neighbors chosen for computing moments was based upon the variance explained for initial embedding of individual patients as well as consistency of nearest neighbors chosen for t-SNE, densMAP, and MNN integration. Full splicing kinetics were recovered using recover_dynamics() for inference of latent time. Next, the velocities were computed using the stochastic model within the velocity() function. The CytoTRACE pseudotime inference was performed using 10000 genes as input as implemented in CellRank (v1.3.1). Driver gene inference for lineage commitment was performed using the above RNA velocity approach and n_top_genes=10000.

Data and Code availability

K562 raw data and fallopian tube counts have been deposited as GSE181544. Raw FASTQ fallopian tube epithelium data will be deposited in dbGaP as controlled access at time of publication. Flow cytometry data will be deposited in FlowRepository and will be made available as above.

Acknowledgements

We would like to acknowledge all members of the Shen and Triche labs as well as the Van Andel Institute Flow Cytometry and Genomics Core facilities. We would like to thank Takara Bio USA for providing resources for protocol development. This research is supported by National Cancer Institute grant 5R37CA230748-04 to HS, philanthropic grants from the Grand Rapids Community Foundation and the Michelle Lunn Hope Foundation to TJT, and startup funds from the Van Andel Institute to TJT. We would like to thank Motion City Soundtrack for the soundtrack to the development of STORM-seq. We would also like to thank the author's kids ('the Coronials') for tolerating parents who constantly put them in front of YouTube to work on this pandemic project.

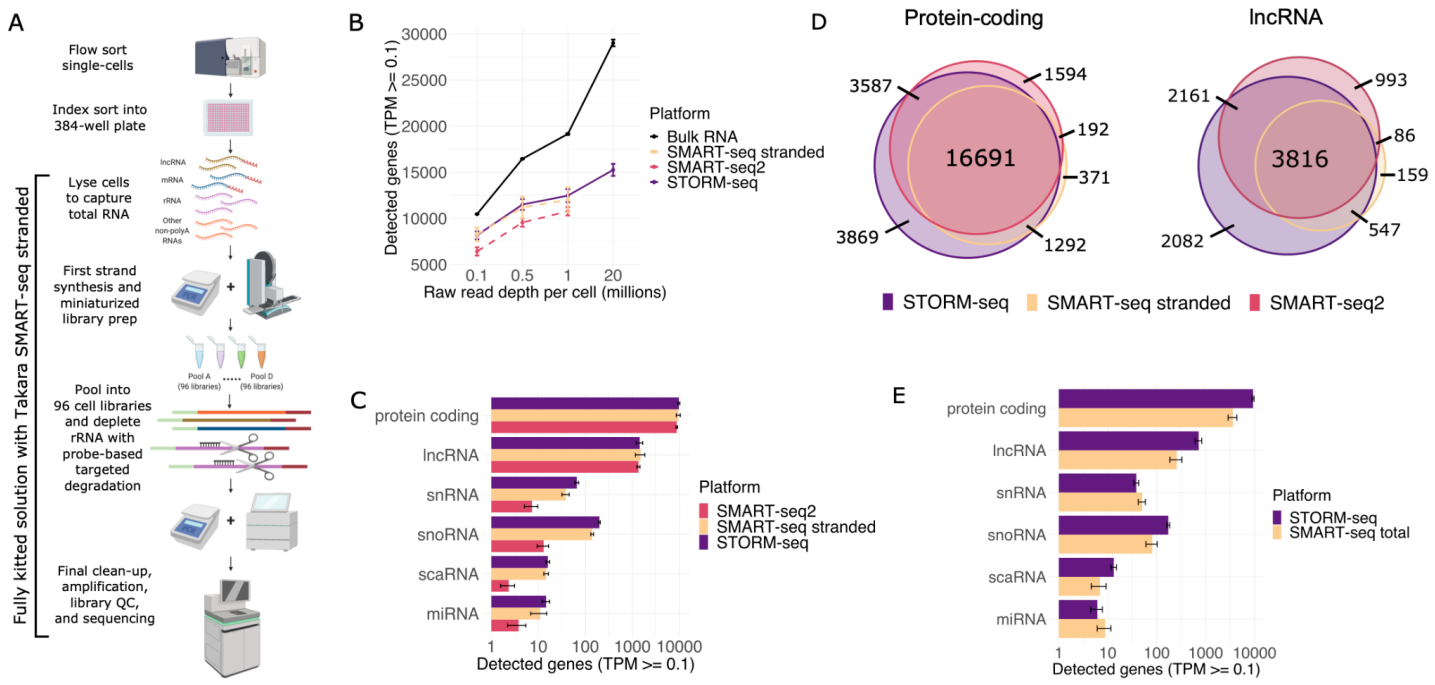

Figure 1. STORM-seq profiles total RNA in single cells.

A) STORM-seq is built upon the Takara SMART-seq stranded protocol with modifications and optimizations to decrease reaction volumes and cost while increasing throughput. B) STORM-seq of K562 (n=70 cells) identifies more genes per/cell across read depths compared to SMART-seq2 (n=65 cells). STORM-seq performs at least as well as the parent, full-volume reaction SMART-seq stranded kit following manufacturer recommendations (n=8 cells). Bulk ribo-reduced total RNA K562 libraries (Kapa hyper prep) are shown as the average of two biological replicates. C) STORM-seq identifies more annotated genes per biotype than either SMART-seq2 or the parent, full-volume reaction SMART-seq stranded kit following manufacturer recommendations. Read depth is 1M raw reads/cell. D) Comparison of STORM-seq, SMART-seq2, and full-volume reaction SMART-seq stranded kit following manufacturer recommendations at 1M raw reads/cell of the overlapping annotated genes in protein-coding and lncRNA space. E) Comparison of STORM-seq (n=120 cells) and SMART-seq total (n=228 cells) in HEK293T cells across annotated gene biotypes. STORM-seq identifies more protein coding and lncRNA species per cell than SMART-seq total. Read depth is 1M raw reads/cell. Annotations used for all gene detection rates was Ensembl 102.

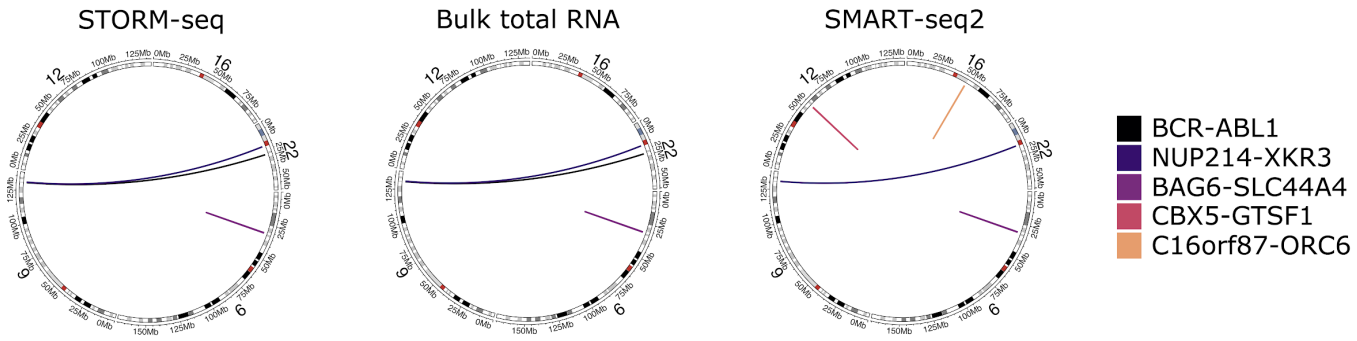

Figure 2. STORM-seq recovers clinically relevant gene fusions in single cells.

STORM-seq identifies gene fusions at this read depth (5M raw reads/cell), similar to bulk ribo-reduced total RNA. STORM-seq identifies *BCR-ABL1* in 85% of single cells (n=70 cells assayed in total). SMART-seq2 did not identify this fusion in any single cells. SMART-seq2 identified *CBX5-GTSF1* (not known to exist in K562 and considered a false positive) and *C16orf87-ORC6*. The latter is found at greater read depths in bulk total RNA and is a known fusion in K562. Gene fusions were called using STAR-fusion in GRCh38. Validated variants are defined as those recovered by deep phased whole-genome reassembly in Zhou et al (2019).

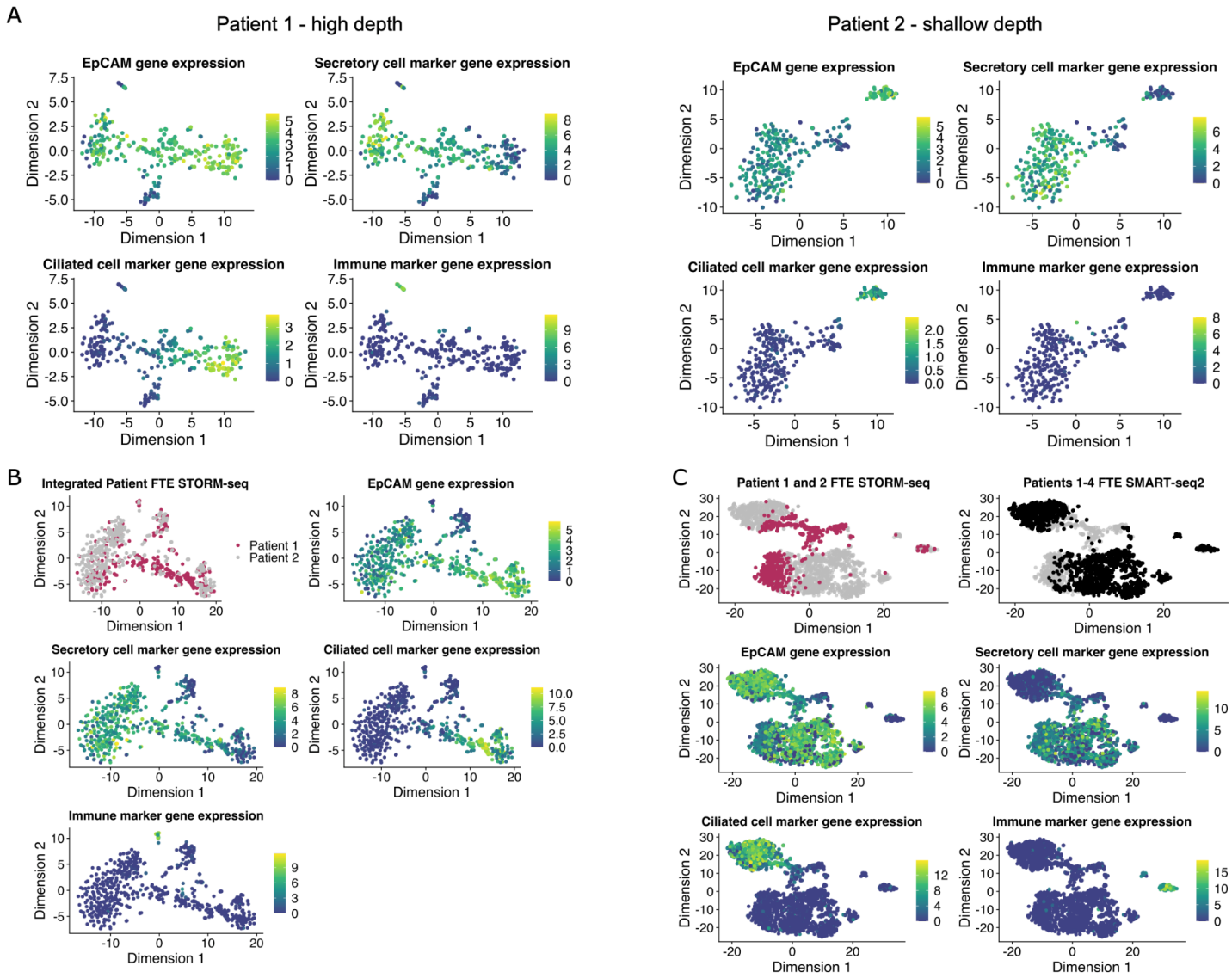

Figure 3. STORM-seq recovers known and intermediate cell types in primary human fallopian tube epithelium across patients and read depths.

A) Patient 1 (deep sequencing, 5-20M reads/cell) and patient 2 (shallow sequencing, 500K-1M reads/cell) primary fallopian tube epithelium (FTE) were compared for recovery of known cell types within the FTE. Log2(TPM+1) is shown. Marker gene expression was summed. Secretory cell marker genes: *PAX8* and *KRT7*. Ciliated cell marker genes: *FOXJ1*, *CCDC17*, and *CCDC78*. Immune cell marker genes: *CD45*, *ITGAX* (CD11c), and *CD14*. B) Integration of patients 1 and 2 via mutual nearest neighbors (MNN) shows that STORM-seq recovers known and intermediate cell types in FTE across patients and read depths. Log2(TPM+1) is shown. C) Integration of STORM-seq with normal FTE from SMART-seq2 (Hu et al. 2019, Cancer Cell) via MNN demonstrates recovery of known cell types in the FTE across protocols. Notably, STORM-seq recovers more intermediate cells than SMART-seq2. All embeddings shown are t-SNE. Gating strategy for cell sorting was singlet, viable (DAPI dim), EpCAM+, CD238a- cells.

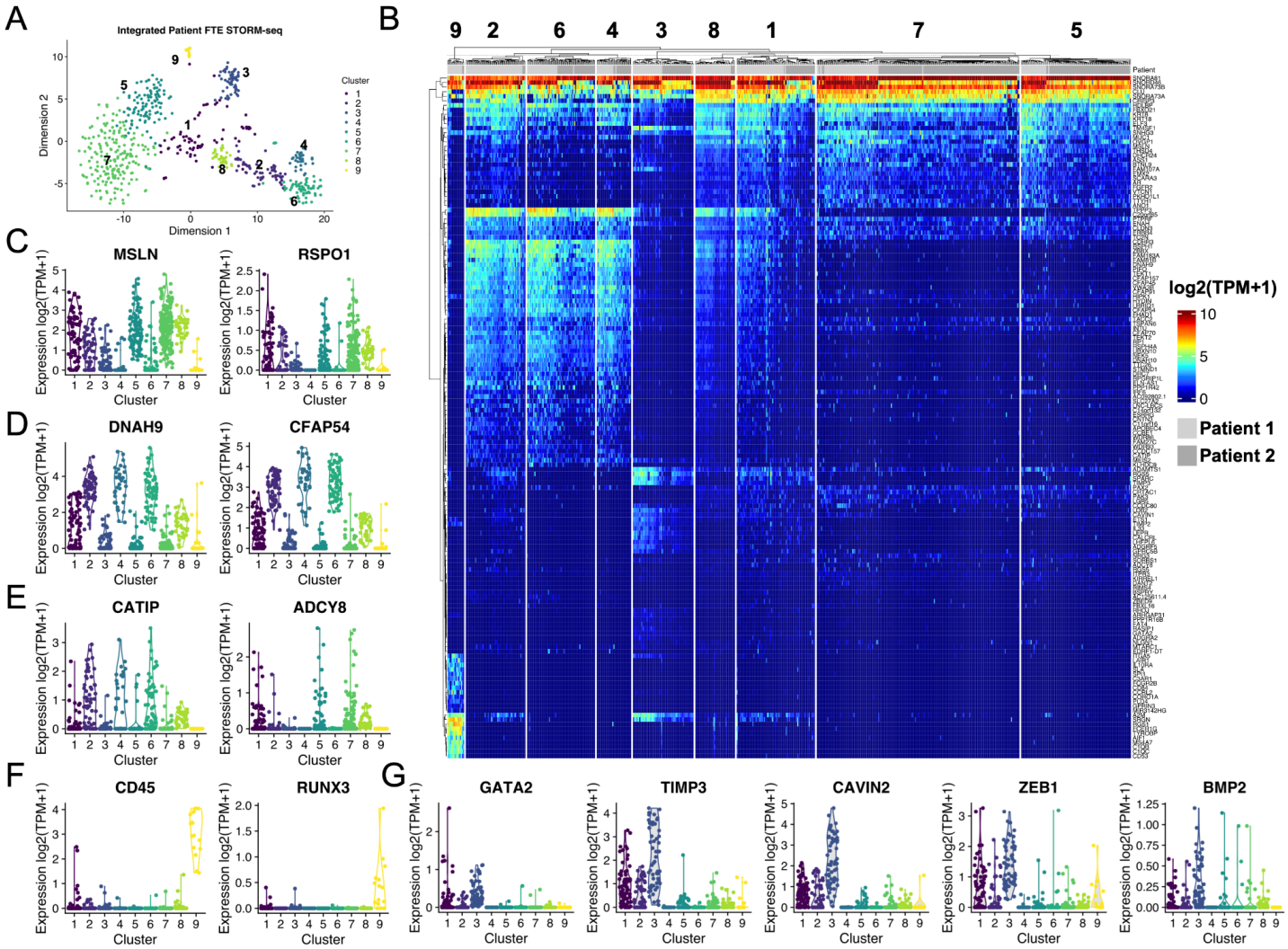

Figure 4. Marker gene detection in the primary FTE identifies novel gene programs for early, intermediate, and terminally differentiated cell types.

A) MNN-integration of patients 1 and 2 identifies distinct clusters of “root-like” (cluster 3), “intermediate” (cluster 8), ciliated intermediate cells (cluster 2), secretory intermediate cells (cluster 1), and multiple subpopulations within secretory (clusters 5 and 7) and ciliated (clusters 4 and 6) terminally differentiated cells. Further, a unique population of EpCAM+, CD45 expressing cells was also identified (cluster 9). B) Heatmap of the top 20 marker genes per cluster. Top marker genes were identified using pairwise Wilcoxon rank sum tests, and keeping significant ($q < 0.05$) genes that were discriminatory for their cluster based on area under the curve (AUC). C) Example marker gene expression for secretory cells (clusters 5 and 7). D) Example marker gene expression for ciliated cells (clusters 4 and 6). E) Example marker gene expression for “intermediate” cells (cluster 8). F) Example marker gene expression for “immune” population (cluster 9). G) Example marker gene expression for cluster 3 (“root-like”, progenitor population). All expression values shown are log2(TPM+1).

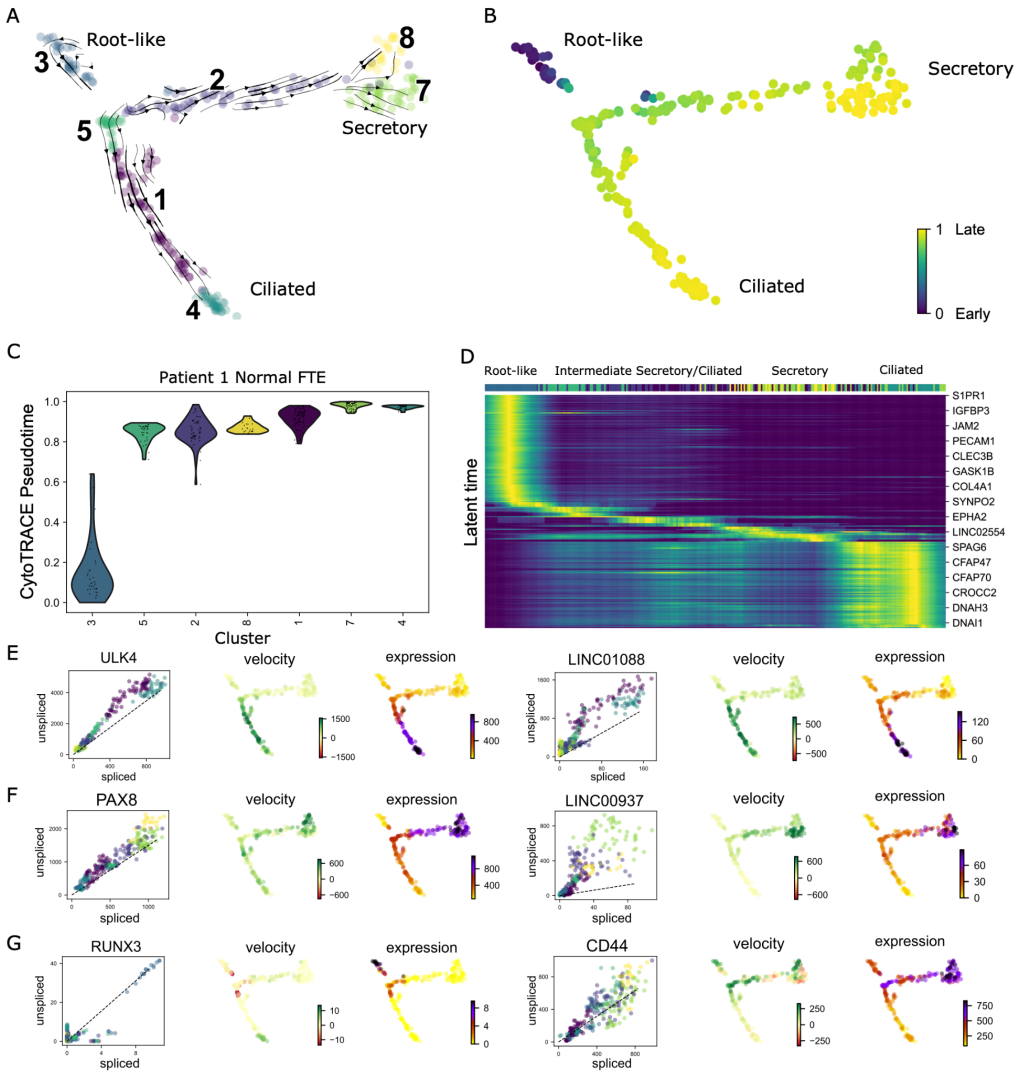

Figure 5. STORM-seq identifies a continuous differentiation trajectory from a bipotent "root-like" cell population in primary human fallopian tube epithelium.

A) RNA velocity supports a continuous differentiation process from "root-like" cells (cluster 3) to an intermediate branch point (cluster 5) to terminally differentiated secretory and ciliated cells (clusters 7-8 and 4, respectively). Thickness and direction of arrows indicate velocity. B) CytoTRACE pseudotime supports a similar differentiation trajectory from early bipotent "root-like" cells towards late, secretory and ciliated cell types. C) CytoTRACE pseudotime organization of cells stratified by cluster. D) Latent time derived from RNA velocity further supports continuous differentiation from a "root-like" population of cells, with key genes associated within each cluster. E) Ciliated and F) secretory cell phase portraits of RNA velocity driver genes indicate *ULK4* (ciliated), *PAX8* (secretory), and long intergenic non-coding RNAs as critical for differentiation trajectories. G) *RUNX3* shows low velocity and does not support a dedifferentiation trajectory. *CD44* is a proposed marker for stem-like cells in human FTE. Colors in the spliced/unspliced phase portrait are colored by cluster as in A). Green shows positive velocity and red negative velocity. Expression is modeled by log transformed gene counts. RNA velocity, CytoTRACE pseudotime, and latent time were calculated using scvelo. densMAP embeddings shown for patient 1.

References

1. Chen, G., Ning, B. & Shi, T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.* **10**, 317 (2019).
2. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
3. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).
4. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
5. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
6. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
7. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0497-0.
8. Hagemann-Jensen, M., Ziegenhain, C. & Sandberg, R. Scalable full-transcript coverage single cell RNA sequencing with Smart-seq3xpress. *bioRxiv* 2021.07.10.451889 (2021) doi:10.1101/2021.07.10.451889.
9. Isakova, A., Neff, N. & Quake, S. R. Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
10. Salmen, F. *et al.* Droplet-based Single-cell Total RNA-seq Reveals Differential Non-Coding Expression and Splicing Patterns during Mouse Development. *bioRxiv* 2021.09.15.460240 (2021) doi:10.1101/2021.09.15.460240.
11. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA

microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).

12. Verboom, K. *et al.* SMARTer single cell total RNA sequencing. *Nucleic Acids Res.* **47**, e93 (2019).
13. Eddy, C. A. & Pauerstein, C. J. Anatomy and physiology of the fallopian tube. *Clin. Obstet. Gynecol.* **23**, 1177–1193 (1980).
14. Ferenczy, A., Richart, R. M., Agate, F. J., Jr, Purkerson, M. L. & Dempsey, E. W. Scanning electron microscopy of the human fallopian tube. *Science* **175**, 783–784 (1972).
15. Paik, D. Y. *et al.* Stem-Like Epithelial Cells Are Concentrated in the Distal End of the Fallopian Tube: A Site for Injury and Serous Cancer Initiation. *Stem Cells* **30**, 2487–2497 (2012).
16. Ulrich, N. D. *et al.* Cellular heterogeneity of human fallopian tubes in normal and hydrosalpinx disease states identified by scRNA-seq. *bioRxiv* 2021.09.16.460628 (2021) doi:10.1101/2021.09.16.460628.
17. Peters, W. M. Nature of 'basal' and 'reserve' cells in oviductal and cervical epithelium in man. *J. Clin. Pathol.* **39**, 306–312 (1986).
18. Hu, Z. *et al.* The Repertoire of Serous Ovarian Cancer Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube Epithelial Cells. *Cancer Cell* **37**, 226–242.e7 (2020).
19. Dinh, H. Q. *et al.* Single-cell transcriptomics identifies gene expression networks driving differentiation and tumorigenesis in the human fallopian tube. *Cell Rep.* **35**, 108978 (2021).
20. Ardighieri, L. *et al.* Characterization of the immune cell repertoire in the normal fallopian tube. *Int. J. Gynecol. Pathol.* **33**, 581–591 (2014).
21. Haas, B. J. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* 120295 (2017) doi:10.1101/120295.
22. Stanoszek, L. M. *et al.* Quality control methods for optimal BCR-ABL1 clinical testing in human

whole blood samples. *J. Mol. Diagn.* **15**, 391–400 (2013).

23. Gottlieb, A. & Tang, Y.-C. PathDSP: Explainable Drug Sensitivity Prediction through Cancer Pathway Enrichment. *Cold Spring Harbor Laboratory* 2020.11.09.374132 (2020) doi:10.1101/2020.11.09.374132.
24. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).
25. Bowen, N. J. *et al.* Emerging roles for PAX8 in ovarian cancer and endosalpingeal development. *Gynecol. Oncol.* **104**, 331–337 (2007).
26. Zhang, S. *et al.* Both fallopian tube and ovarian surface epithelium are cells-of-origin for high-grade serous ovarian carcinoma. *Nat. Commun.* **10**, 5367 (2019).
27. Sletten, A. C. *et al.* Loss of SNORA73 reprograms cellular metabolism and protects against steatohepatitis. *Nat. Commun.* **12**, 5214 (2021).
28. Howell, J. J. & Manning, B. D. mTOR couples cellular nutrient sensing to organismal metabolic homeostasis. *Trends Endocrinol. Metab.* **22**, 94–102 (2011).
29. Leese, H. J., Tay, J. I., Reischl, J. & Downing, S. J. Formation of Fallopian tubal fluid: role of a neglected epithelium. *Reproduction* **121**, 339–346 (2001).
30. Olsson, R. The relationship between ciliary rootlets and other cell structures. *J. Cell Biol.* **15**, 596–599 (1962).
31. Morse, D. M., Smullen, J. L. & Davis, C. W. Differential effects of UTP, ATP, and adenosine on ciliary activity of human nasal epithelial cells. *Am. J. Physiol. Cell Physiol.* **280**, C1485–97 (2001).
32. Lin, W., Qiao, C., Hu, J., Wei, Q. & Xu, T. Conserved role of ATP synthase in mammalian cilia. *Exp. Cell Res.* **401**, 112520 (2021).
33. Erzberger, J. P. & Berger, J. M. Evolutionary relationships and structural mechanisms of AAA+

proteins. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 93–114 (2006).

34. King, S. M. & Sale, W. S. Fifty years of microtubule sliding in cilia. *Mol. Biol. Cell* **29**, 698–701 (2018).
35. Raoux, M. *et al.* Multilevel control of glucose homeostasis by adenylyl cyclase 8. *Diabetologia* **58**, 749–757 (2015).
36. Roger, B. *et al.* Adenylyl cyclase 8 is central to glucagon-like peptide 1 signalling and effects of chronically elevated glucose in rat and human pancreatic beta cells. *Diabetologia* **54**, 390–402 (2011).
37. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
38. Bergen, V., Lange, M., Peidli, S., Alexander Wolf, F. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *bioRxiv* 820936 (2019) doi:10.1101/820936.
39. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405–411 (2020).
40. Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* **47**, e66 (2019).
41. Hayashi, T. *et al.* Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat. Commun.* **9**, 619 (2018).
42. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).
43. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26**, 1271–1279 (2020).
44. Chen, G., Deng, C. & Li, Y.-P. TGF-β and BMP signaling in osteoblast differentiation and bone formation. *Int. J. Biol. Sci.* **8**, 272–288 (2012).

45. Choi, Y.-J. *et al.* Identifying an ovarian cancer cell hierarchy regulated by bone morphogenetic protein 2. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6882–8 (2015).

46. May, G. *et al.* Dynamic analysis of gene expression and genome-wide transcription factor binding during lineage specification of multipotent progenitors. *Cell Stem Cell* **13**, 754–768 (2013).

47. Wheat, J. C. *et al.* Single-molecule imaging of transcription dynamics in somatic stem cells. *Nature* (2020) doi:10.1038/s41586-020-2432-4.

48. Dobrzycki, T. *et al.* Deletion of a conserved Gata2 enhancer impairs haemogenic endothelium programming and adult Zebrafish haematopoiesis. *Commun Biol* **3**, 71 (2020).

49. Bluteau, O. *et al.* A landscape of germ line mutations in a cohort of inherited bone marrow failure patients. *Blood* **131**, 717–732 (2018).

50. Liu, M. *et al.* Ulk4 Is Essential for Ciliogenesis and CSF Flow. *J. Neurosci.* **36**, 7589–7600 (2016).

51. Ai, H. *et al.* The down-regulation of long non-coding RNA LINC01088 is associated with the poor prognosis of epithelial ovarian cancer patients. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 5836–5841 (2018).

52. Zhang, W. *et al.* LINC01088 inhibits tumorigenesis of ovarian epithelial cells by targeting miR-24-1-5p. *Sci. Rep.* **8**, 2876 (2018).

53. Shi, K. *et al.* PAX8 regulon in human ovarian cancer links lineage dependency with epigenetic vulnerability to HDAC inhibitors. *Elife* **8**, (2019).

54. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Cold Spring Harbor Laboratory* 2020.10.19.345983 (2020) doi:10.1101/2020.10.19.345983.

55. Berrens, R. V. *et al.* Locus-specific expression of transposable elements in single cells with CELLO-seq. *Nat. Biotechnol.* 1–9 (2021) doi:10.1038/s41587-021-01093-1.

56. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**,

137–145 (2019).