Applications of Next-Generation Sequencing to Rare Disease

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Catherine Ann Alsager Lee

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jakub Tolar

July, 2018

Acknowledgements

First and foremost, I would like to acknowledge the recessive dystrophic epidermolysis bullosa patients and their siblings as well as the childhood cerebral adrenoleukodystrophy patients and other healthy individuals who donated cells and/or tissue to make these studies possible. Their willingness to participate in basic research makes possible new treatments not only for themselves but for others suffering from these devastating diseases.

I am immensely thankful to my advisor, Dr. Jakub Tolar, for being a constant inspiration, a wealth of knowledge, and a thoughtful mentor who allowed me to explore several very different projects during my time in his lab. I truly appreciate all he has contributed with his time, ideas, and funding to make my Ph.D. experience productive and stimulating. The enthusiasm he has for translating basic research into treatments for patients kept me motivated during arduous experiments and setbacks in the lab.

For this dissertation, I would like to thank my reading committee members: Drs. Melissa Gardner, Rui Kuang, and Jakub for their time and insightful comments. I would like to thank the three of them and the other members of my oral prelim and final defense committees: Drs. Ran Blekhman, Rita Perlingeiro, and Dan Voytas for all of their support and advice throughout my years as a graduate student.

I am thankful to those with whom I have co-authored manuscripts and worked closely with for my various projects. Without the knowledge and assistance of all of these individuals, completion of these projects would not have been possible.

Regarding the single-cell work, from the Computer Science Department, I thank Drs. Huanan Zhang and Rui Kuang for helpful discussions and computational expertise and for sticking it out with me through the long process of publishing the single-cell work. Their hard work alongside Zhuliu Li and Raphael Petegrosso allowed us to formalize and test the clustering algorithm we developed. I am also grateful to John Garbe at the Minnesota Supercomputing Institute for creating a pipeline for single-cell RNA-seq data analysis and to Cindy Eide for helping with the flow cytometry experiments.

Regarding the BMEC project, from the Chemical Engineering and Materials Science Department, I thank Dr. Samira Azarin for sharing her expertise in directed differentiations and for her clear-headed logic, unwavering guidance, and insightful conversations regarding the barrier studies. I would also like to thank Dr. Frank Bates for his contributions from the field of polymer science and for the use of a polymer that was custom-designed in his lab. I am immensely indebted to Hannah Seo for all of her help with the differentiations, numerous experiments, and polymer testing. I appreciated her camaraderie as well as her incredible diligence and attention to detail. She made the late nights and long hours we spent on this project a much more enjoyable experience. Additionally, Dr. Anibal Armien and Dean Muldoon at the Veterinary Diagnostic Lab prepared the EM samples. Dr. Juan Abrahante from the University of Minnesota Informatics Institute assisted with RNA-seq data analysis. LeAnn Oseth and Paula Haffner from the University of Minnesota Cytogenomics Laboratory did the genetic fingerprinting and karyotyping of the cell lines. Dr. Mihee Kim helped in preparing the polymer samples. Abby Silbaugh, Hope Leslie, and Faith Leslie assisted with the frayed junction analysis. Valencia Owens assisted with the iPSC and iBMEC cell culture and Anna Xue assisted with the sodium fluorescein permeability assay and iPSC and iBMEC cell culture.

Regarding the mosaic project, I thank Dr. Michael Vanden Oever for initiating the project, teaching me about miRNA and transfections, and passing the project on to me. I would also like to thank Dr. Beau Webber and Chris Lees for initiating

the PacBio experiments and Kevin Silverstein at the Minnesota Supercomputing Institute for help with the PacBio data analysis. Amber McElroy assisted with the transfections and Beth Thompson performed the flow cytometry.

For all of these projects, I would like to thank the University of Minnesota Genomics Center for quality control of the NGS libraries used in the studies as well as Nancy Morgan and Susan Julson for administrative support.

Many members of the Tolar and Kuang Labs have made immensely positively contributions to my personal and professional time at the University of Minnesota. The lab groups have been a source of friendships as well as advice and collaboration. I am especially grateful for the group who worked with me at the Stem Cell Institute: Weili Chen, Emily Ward, Beth Thompson, Mike Pickett-Leonard, Mike Vanden Oever, Valenica Owens, Allie Keith, Ilona Rousalova, and Kirk Twaroski as well as numerous rotation and summer students. I am also grateful for the support from our group at the Cancer Center: Cindy Eide, Lily Xia, Megan Riddle, Wendy Matthews, Ron McElmurry, Madison Mack, and Chris Lees as well as the group in Keller Hall: Huanan Zhang, Wei Zhang, Raphael Petegrosso, Nissi Paidimukkala, and David Roe.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. During my third and fourth years, I was funded by the National Institutes of Health through a T32 Stem Cell Biology training grant. I was also honored to receive a Society for Investigative Dermatology Eugene M. Farber Travel Award for Young Investigators and a Center for Genome Engineering Travel Award that funded travel to several important conferences for oral and poster presentations.

My time at the University of Minnesota was made all the more enjoyable in large part due to the many friends I made through the MCSB program. The time spent with them talking research or just commiserating at Stubs made the good times all the better and the hard times bearable. I look back fondly on the weekend outings and road trips we managed to fit in and, of course, the yearly trip to the University of Minnesota Biological Station at Lake Itasca.

Dedication

This dissertation is dedicated to my family. To my parents, for their ceaseless love, encouragement, and support of me during the degree-seeking process. To my brother, Christopher, who attended medical school at the University of Minnesota during the first four years of my Ph.D. And most of all to my loving, wonderful, and patient husband, Tristan, whose support during the most challenging times was instrumental to my success. Thank you.

Abstract

Since the discovery of the structure of DNA in 1953, researchers and clinicians have been painstakingly paving the way for the use of genetic information in the treatment of disease. In order for this to be possible, specific genetic targets must be identified. For this dissertation, I use next generation, single-cell, and third generation RNA-sequencing techniques to identify markers of genetic heterogeneity and potential therapeutic targets in the rare diseases recessive dystrophic epidermolysis bullosa (RDEB) and cerebral childhood adrenoleukodystrophy (ccALD).

RDEB is a an inherited blistering disorder caused by mutations in the key structural skin protein, type VII collagen (C7). It can partially treated by hematopoietic stem cell transplant (HSCT), however, how blistered RDEB skin signals to donor cells is unknown. In Chapter 2, to identify potential signals, I performed single-cell RNA-seq (scRNA-seq) on patient fibroblasts and implemented a variance-driven multitask clustering (scVDMC), that utilizes multiple single-cell populations from biological replicates or different samples. scVDMC clusters single cells in multiple scRNA-seq experiments of similar cell types and markers but varying expression patterns such that the scRNA-seq data are better integrated than typical pooled analyses which only increase the sample size. By controlling the variance among the cell clusters within each dataset and across all the datasets, scVDMC detects cell sub-populations in each individual experiment with shared cell-type markers but varying cluster centers among all the experiments. scVDMC was then applied to two previously published scRNA-seq datasets with several replicates and one large-scale Drop-seq dataset on three patient samples. scVDMC more accurately detected cell populations and known cell markers than pooled clustering and other recently proposed scRNA-seq clustering methods. When applied to the scRNA-seq RDEB patient fibroblast data, scVDMC revealed several new cell types and unknown markers that I validated by flow cytometry.

ccALD is caused by mutations in the *ABCD1* gene and manifests in early childhood with neuropathological symptoms and hyper-pigmentation, culminating in massive breakdown of the blood-brain barrier (BBB) and death if HSCT is not performed at an early stage. It is difficult to model the BBB of this disease as primary cells do not recapitulate the barrier in culture and the mouse model shows incomplete penetrance. In Chapter 3, I model the blood-brain barrier of ccALD patients and wild-type (WT) controls using directed differentiation of induced pluripotent stem cells (iPSCs) into induced brain microvascular endothelial cells (iBMECs). Immunocytochemistry and PCR confirmed characteristic expression of brain microvascular endothelial cell (BMEC) markers. Barrier properties of iBMECs were measured via trans-endothelial electrical resistance (TEER), sodium fluorescein permeability, and frayed junction analysis. Electron microscopy and RNA-seq were used to further characterize disease-specific differences. Oil-Red-O staining was used to quantify differences in lipid accumulation. To evaluate whether treatment with block copolymers of poly(ethylene oxide) and poly(propylene oxide) (PEO-PPO) could mitigate defective properties, ccALD-iBMECs were treated with PEO-PPO block copolymers and their barrier properties and lipid accumulation levels were quantified. iBMECs from patients with ccALD had significantly decreased TEER ($2592 \pm 110\ \Omega \cdot cm^2$) compared to WT controls ($5001 \pm 172\ \Omega \cdot cm^2$). They also accumulated lipid droplets to a greater extent than WT-iBMECs. Upon treatment with a PEO-PPO diblock copolymer during the differentiation process, an increase in TEER and a reduction in lipid accumulation were observed for the polymer treated ccALD-iBMECs compared to untreated controls. The finding that BBB integrity is decreased in ccALD and can be rescued with block copolymers opens the door for the discovery of BBB-specific molecular markers that can indicate the onset of ccALD and has therapeutic implications for preventing the conversion to ccALD.

Revertant mosaicism in RDEB patients is seen as patches of skin that have never blistered. At the molecular level, these patches of skin contain detectable amounts of C7, indicating that a reversion of the disease-causing mutation has

occurred at the DNA level. One of the limited treatment options available for treating RDEB is the use of C7 expressing stem cells or differentiated skin cells to replace C7 at the dermal-epidermal junction and restore the overall integrity of the skin architecture. However, this typically requires the use of gene therapy or allogeneic cells, which can be costly and cause adverse reactions in the recipient. Mosaic cells could potentially be used for these purposes, however, isolating and purifying them has proven difficult. In Chapter 4, I describe a method utilizing synthetic micro RNA (miR) switches, whereby differences in endogenous miR activity are exploited to purify mosaic cells in culture, which may be useful in generating pure populations of mosaic cells that can then be used in future clinical applications. Chapter 5 of this dissertation uses third generation or long read sequencing to look more closely at the underlying genetic event resulting in mosaic expression in one particular RDEB patient.

These studies identify genetic heterogeneity in cell types relevant to the respective rare diseases being examined and give support to developing precision medicine techniques to treat these rare diseases.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human genetics has progressed from a nearly nonexistent state to an observational science and in recent years is finally blossoming into an interventional science [1]. At the forefront of this has been the sequencing of the human genome and the massive parallelization of sequencing that has resulted in vast amounts of genetic information readily and affordably available. Once largely guesswork, the ability to identify and diagnose common genetic diseases (including cancer) is becoming routine [1]. However, the work of geneticists and genome engineers does not end here. Instead, more complete realization of the technology has allowed researchers to use these and other advances in technology to increase the diagnostic search space by building increasingly complex models of genetic, transcriptomic, and proteomic interactions. Very recently, this search space has extended further into single-cell sequencing and has also begun to include information across cell types, individuals, as well as across populations (shown for ccALD in Figure 1.1 and RDEb in Figure 1.2).

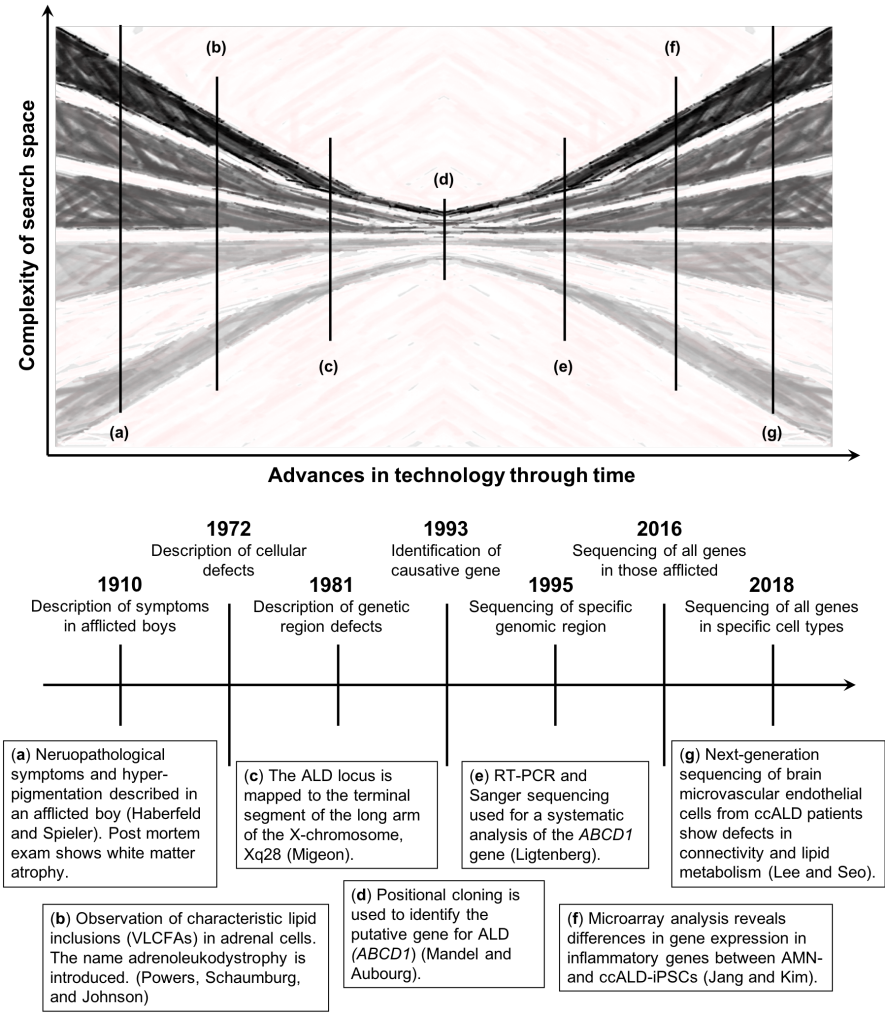


Figure 1.1: **Complexity and scope of identifying and treating ccALD has changed over time as a function of the technology available.** Advances in technology initially converged on the discovery of disease causing genes but with more individuals being diagnosed the technology can now be used to look at factors such as modifier genes and gene expression in specific cell types. Abbreviations: VLCFAs: very long chain fatty acids; ALD: adrenoleukodystrophy; RT-PCR: real-time polymerase chain reaction; AMN: adrenomyeloneuropathy; ccALD: cerebral childhood adrenoleukodystrophy; iPSC: induced pluripotent stem cell.

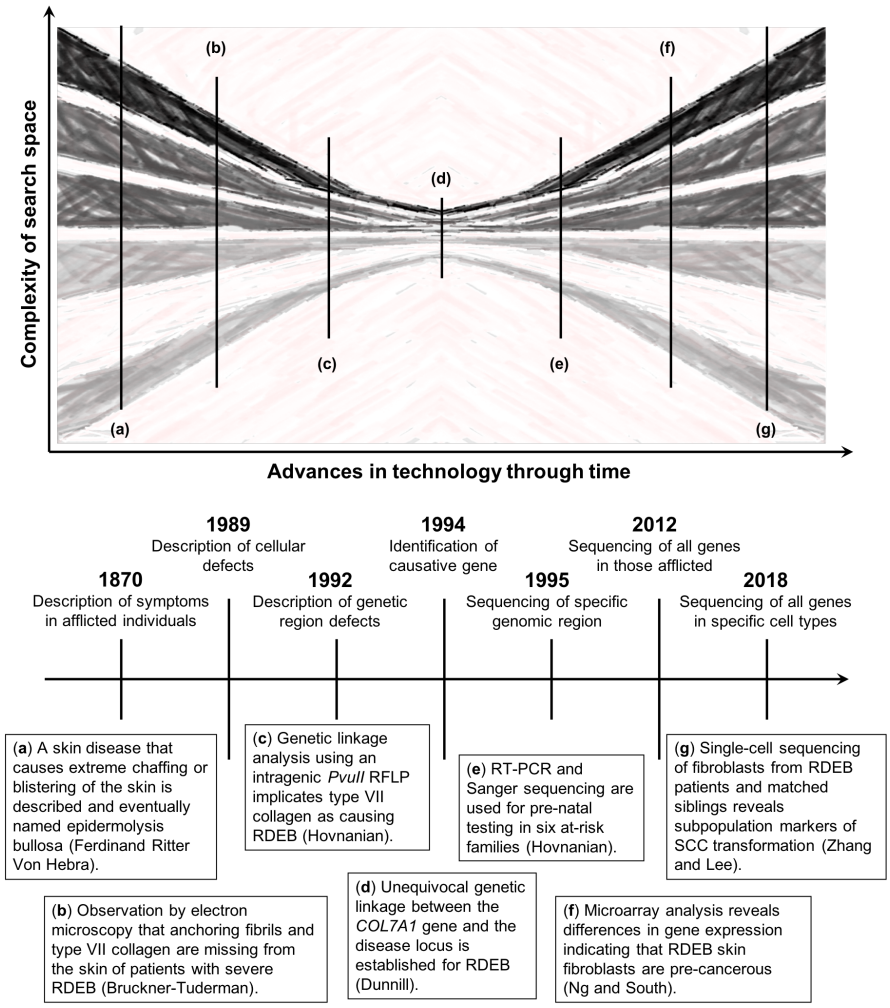


Figure 1.2: **Complexity and scope of identifying and treating RDEB has changed over time as a function of the technology available.** Advances in technology initially converged on the discovery of disease causing genes but with more individuals being diagnosed the technology can now be used to look at additional factors and gene expression in specific cell types. Abbreviations: RDEB: recessive dystrophic epidermolysis bullosa; RFLP: restriction fragment length polymorphism; SCC: squamous cell carcinoma.

The work described in my dissertation is three-fold. First, I have implemented a multitask clustering algorithm to identify specific subpopulations of cells that are involved in disease; in this case, fibroblasts subpopulations from the skin of recessive dystrophic epidermolysis bullosa (RDEB) patients. I used single-cell RNA-sequencing (scRNA-seq) to determine the phenotype (gene expression) of the cells and to find a minority population of cells that contributed to progression of the disease. Similarly, for another rare disease, cerebral childhood adrenoleukodystrophy (ccALD), while most efforts have been directed towards the immune system and how it opens the blood-brain barrier (BBB) in ccALD, I focused on the barrier itself and targeted the particular cell type, brain microvascular endothelial cells (BMECs), that constitute the BBB. Using trans-endothelial electrical resistance (TEER), I found differences in barrier integrity in BMECs from ccALD patients compared to healthy individuals. Working with a chemical engineering and materials science group, we identified a polymer that was able to mitigate this defect. I went further by performing RNA-seq on these same cell types and found differences in lipid metabolism pathways. This strategy of hunting for genetic markers in the cell type of interest was reapplied for the third part of my dissertation, in which I sought to use genetic heterogeneity to separate non-diseased from diseased cells. The genetic markers in this case were microRNAs (miRs), that are differentially expressed between mosaic (non-diseased) and blistered (diseased) fibroblasts in a case of mosaicism, an enigmatic event where genetic reversion results in self-correction of the disease-causing mutation at the DNA level. The goal was to separate the self-corrected fibroblasts using miR-switches, a technique developed by Shinya Yamanaka, and select for this minority population of naturally gene-corrected cells to use in downstream gene therapy applications. Ongoing work for the third part of my dissertation involves taking the project into the third generation of sequencing and using long read sequencing to look in detail at the reversion event occurring in a particular RDEB patient.

1.1 Sequencing

1.1.1 First-generation sequencing

Watson and Crick famously solved the three-dimensional structure of DNA in 1953, working from crystallographic data produced by Rosalind Franklin and Maurice Wilkins [2, 3], which contributed to a conceptual framework for both DNA replication and how proteins are encoded by nucleic acids. The ability to 'read' or sequence DNA, however, did not follow for some time [4]. Early nucleic acid sequencing techniques borrowed heavily from analytical chemistry and were far from efficient [4]. It wasn't until 1977 when Frederick Sanger's 'chain termination' or 'dideoxy technique improved the accuracy and ease of use for sequencing DNA that what is now known as the 'first-generation' DNA sequencing was able to take off [4]. First-generation DNA sequencers produced reads slightly less than one kilobase. To analyze longer fragments, shotgun sequencing was used, where overlapping DNA fragments are cloned and sequenced separately, and then assembled into one long contiguous sequence (or 'contig') *in silico* [5, 6, 4].

Beginning in the mid-1980s, and for the following two decades, the primary approach to gene discovery was a combination of linkage analysis, positional cloning, and sequencing of candidate or regionally selected genes, most of which was hypothesis driven. Researchers became able to identify mutations responsible for Mendelian disorders/monogenic diseases using linked markers starting in the late 1980s.

During the end of this period, the development of polymerase chain reaction (PCR), the use of better polymerases, and the development of newer dideoxy sequencers enabled the Human Genome Project to be completed years ahead of schedule [7, 8, 4]. Completion of the human genome project in 2003 revolutionized the study of genetic disorders [1] and also ushered in a new era of sequencing, in which the human genome could be used as a reference sequence.

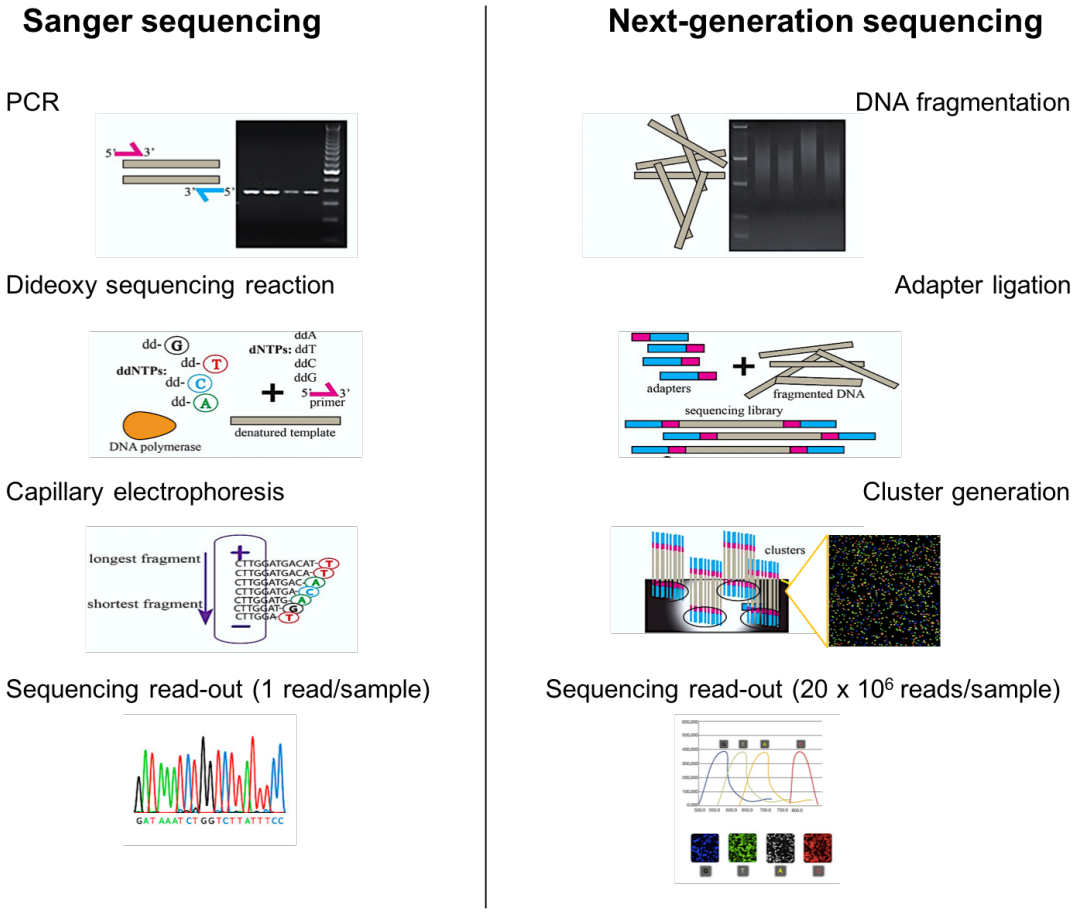


Figure 1.3: **Comparison of Sanger sequencing and next-generation sequencing.** The steps involved in Sanger sequencing (left) generate one read per sample while the mass parallelization approach of next-generation sequencing (right) can generate 20 x 10^6 reads per sample (average for a relatively low throughput instrument such as Illumina's MiSeq or MiniSeq). Adapted from Jaya Punetha and Eric P. Hoffman *Circ Cardiovasc Genet.* 2013;6:427-434 and Illumina.

1.1.2 Next-Generation Sequencing

In the mid-2000s, the next breakthrough that ushered in the second generation or 'next-generation' of sequencing was pyrosequencing [9]. This was first pioneered by 454 Life Sciences and later improved with the use of flowcells by Solexa (subsequently acquired by Illumina) [10]. Sequencing-by-synthesis using fluorescent 'reversible-terminator' dNTPs, which cannot immediately bind further nucleotides as the fluorophore occupies the 3' hydroxyl position. This must be cleaved away before polymerization can continue, allowing sequencing to occur in a synchronous manner, greatly increasing the efficiency [11]. These advances in nucleotide sequencing technology spurred the 'genomics revolution' which lowered the cost of sequencing by increasing the capability of DNA sequencers at a rate that has outpaced Moore's law: the complexity of microchips (measured by number of transistors per unit cost) doubles approximately every two years, while sequencing capabilities between 2004 and 2010 doubled every five months [12]. See Figure 1.3) for a comparison of Sanger and NGS sequencing.

In addition to the increased efficiency, mass parallelization of sequencing reactions using the bridge amplification method developed by Solexa had the added advantage of being paired end (PE) data. In this way, the sequences at both ends of the cluster are recorded [4]. PE sequencing greatly increases the amount of information produced for each read and also improves the accuracy of mapping the reads to reference genomes. Additionally, it facilitates the detection of spliced exons DNA rearrangements.

The introduction of next generation sequencing (NGS) strategies to identify genes associated with disease, primarily based on whole exome sequencing (WES), in 2009 accelerated the pace of causal gene discovery by enabling hypothesis-free approaches. Since 2010, causal mutations for monogenic disorders have been described for nearly 2000 additional genes [13]. Today, WES is routinely used as the primary technological approach to discovering disease-gene associations. Its favor over whole genome sequencing (WGS) has primarily been due to its significantly lower cost and that the majority of pathogenic variants continue to be

within the protein-coding portion of the genome [13]. Such technological advances have generated an unprecedented wealth of new information on underlying defects, mechanisms, and therapeutic targets of rare diseases in the past decade [14].

scRNA-seq enables detailed profiling of heterogeneous cell populations and can be used to reveal lineage relationships or discover new cell types. Growing understanding of single-cell biology has increased at a rapid pace in the last five years, and single-cell systems biology is a new approach that seeks to elucidate the role of single-cells in tissue and organ function and in disease pathogenesis [15]. This represents a novel approach to identify and validate disease-specific markers. Translating the results from single-cell studies into clinical practice remains a challenge but is worthwhile as it could inform patient response to therapies (or lack thereof) [15].

1.1.3 Single-cell RNA-sequencing

In the literature, there has been little effort directed towards developing computational methods for cross-population transcriptome analysis of multiple single-cell populations. The cross-cell-population clustering problem is different from the traditional clustering problem because single-cell populations can be collected from different patients, different samples of a tissue, or different experimental replicates. The accompanying biological and technical variation tends to dominate the signals for clustering the pooled single cells from the multiple populations. In Chapter 2, I detail a multitask clustering method that was developed to address the cross-population clustering problem. The method simultaneously clusters each individual cell population and controls variance among the cell-type cluster centers within each cell population and across the cell populations. This multitask clustering method significantly improves clustering accuracy and marker discovery as demonstrated with three public scRNA-seq datasets and an in-house Recessive Dystrophic Epidermolysis Bullosa (RDEB) dataset. The results of this work make it evident that multitask clustering is a promising new approach for

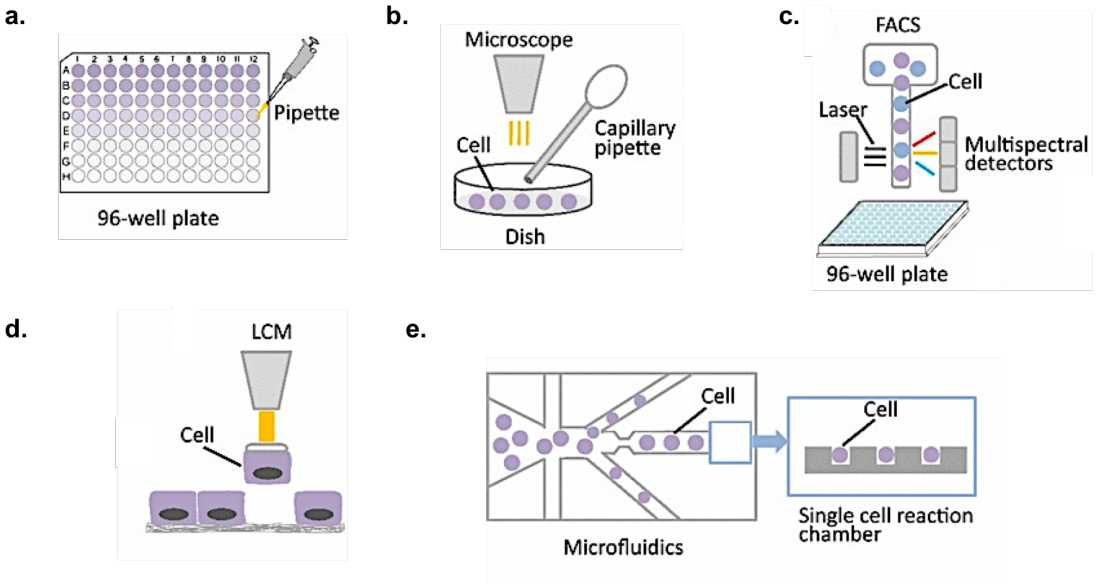


Figure 1.4: **Current methods for single-cell isolation.** (a) Serial dilution. (b) Mechanical micro-manipulation. (c) Laser capture micro-dissection (LCM). (d) Fluorescence activated cell sorting (FACS). (e) Microfluidics. Adapted from Jian Wang and Yuanlin Song *Clin Transl Med.* 2017; 6:10.

cross-population analysis of scRNA-seq data.

1.1.4 Third-generation sequencing

While the division between second- and third-generations sequencing is not as clearly delineated as first- and second-generation sequencing, the most widely used third-generation sequencing technology today is the single-molecule real-time (SMRT) platform from Pacific Biosciences (PacBio) [16]. SMRT sequencing, while not as high throughput as NGS, allows sequencing of non-amplified DNA, thus circumventing issues with bias and errors during PCR amplification [4] (Figure 1.5). PacBio sequencing is also capable of producing very long reads (up to and exceeding 10 kb in length) which is highly useful for *de novo* genome assembly as well as resolution of repetitive stretches of DNA [4]. In Chapter 5, I use PacBio sequencing to investigate the cause of mosaicism in a patient with recessive dystrophic epidermolysis bullosa.

1.2 Rare disease

Rare diseases, though individually rare, are collectively common. A rare disease is defined as one that affects fewer than 200,000 people in the US [17] or less than 1 in 2,000 people in Europe [18]. A substantive number of rare diseases are due to altered functions of single genes. Cumulatively, these rare genetic diseases, also termed Mendelian or monogenic diseases, affect at least 1 in 50 individuals in the European-derived general population [19, 13].

With many common and some rare diseases, next-generation sequencing (NGS) has been instrumental in identifying disease-causing genes and for deciphering gene and protein signatures [14]. The former is extremely important for rare diseases because a timely, molecularly confirmed diagnosis for children and adults with rare genetic disease drastically shortens the "diagnostic odyssey" that so often accompanies rare disease. In addition, disease management and outcomes

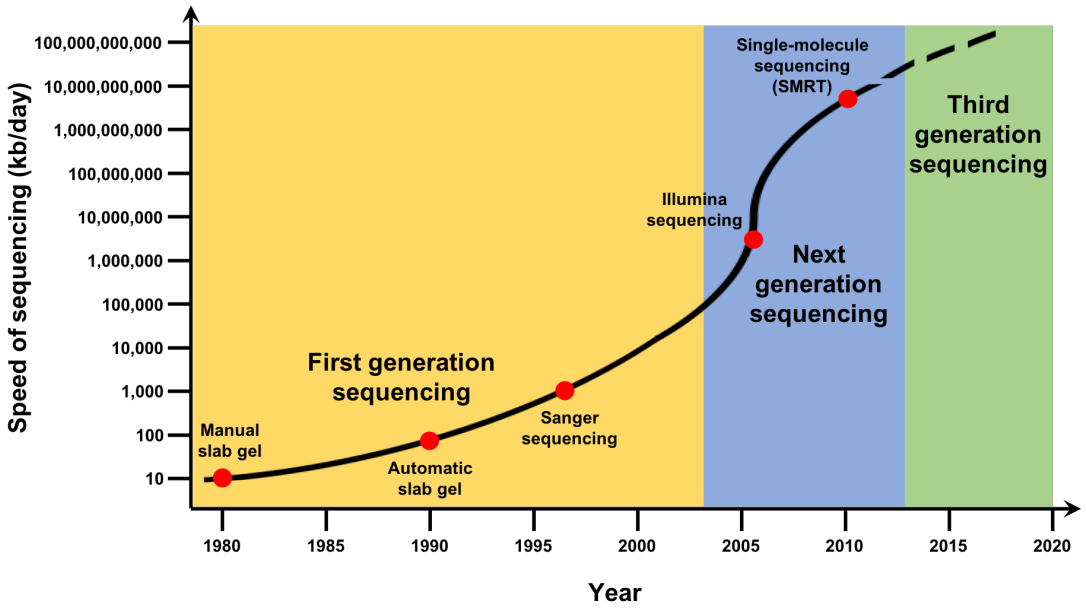


Figure 1.5: **The speed of DNA sequencing has increased more than exponentially in recent decades.** Early techniques developed in the 1980s paved the way for later techniques that greatly increased the speed at which DNA could be sequenced. With the introduction of longer read lengths with third generation sequencing and the decreased cost/increased sensitivity, scientists can now go back and re-sequence genomes that have already been sequenced to achieve a higher level of accuracy. Adapted from *Genome Research Limited.*

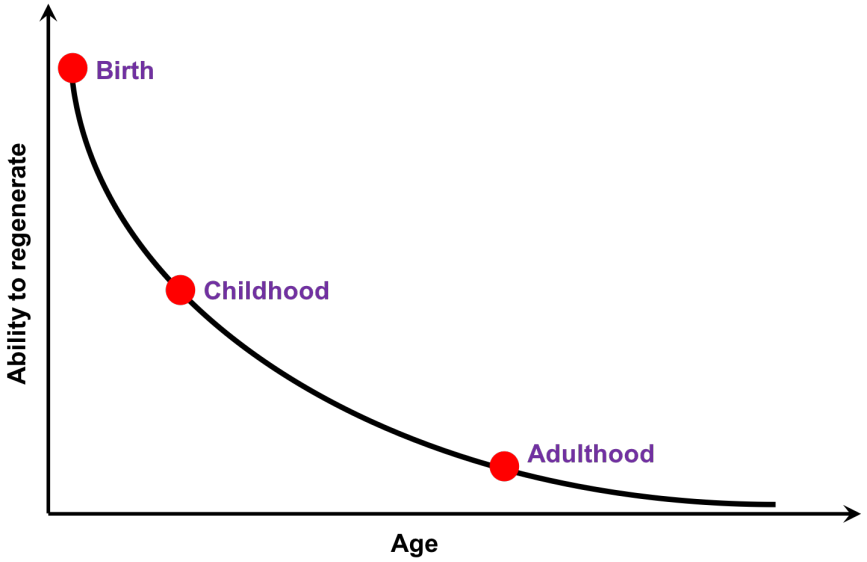


Figure 1.6: **The same genotype results in different clinical presentations over time.** As patients, age, disease burden increases and more reactive changes (immunological and other) occur. Typically, there is less capacity to respond to treatment. That different treatments are relevant to patients of different ages with the same genotype is obvious but often overlooked. Adapted from Tolar, J. *Clin Pharmacol Ther.* 2018 Jan 30.

are improved and genetic counselors are able to assess recurrence risk regarding reproductive choices. In the average clinical genetics setting, the current diagnostic rate is approximately 50%, but for those who do not receive a molecular diagnosis after the initial genetics evaluation, that rate is much lower. Diagnostic success for these more challenging affected individuals depends to a large extent on progress in the discovery of genes associated with, and mechanisms underlying, rare diseases [13]. Additionally, despite their often chronic and progressive nature, long-term complications can be lessened or delayed for some rare diseases if they are diagnosed early (e.g. via newborn screening) and optimally managed by standard treatment or targeted therapies (Figure 1.6). A definitive molecular diagnosis can obviate the need for further diagnostic investigations, facilitate appropriate access to health care resources, reduce prognostic uncertainty, provide accurate recurrence risk counseling, foster reproductive choices in affected families, and impart psychosocial benefits to the patient and their family [13].

The two rare diseases pertinent to this dissertation are recessive dystrophic epidermolysis bullosa (RDEB) and cerebral childhood adrenoleukodystrophy (ccALD). Both of these diseases are monogenic diseases in which the causative gene is know. Both can be treated (at least partially) by HSCT. For both, however, questions remain regarding disease progression and the mechanism of action of donor cells.

1.2.1 Recessive dystrophic epidermolysis bullosa (RDEB)

Recessive Dystrophic Epidermolysis Bullosa (RDEB) is an inherited blistering disorder caused by loss-of-function mutations in the *COL7A1* gene that codes for type VII collagen (C7) [20]. C7 forms the anchoring fibrils that attach the epidermis to the dermis [21]. When C7 is missing, the skin becomes extremely fragile, eroding at the slightest touch. From birth, patients with this disease must undergo intensive bandaging and daily wound care. Mutilating scarring is a hallmark of this disease with patients at high risk for anemia and infections. They are also susceptible to a highly aggressive form of squamous cell carcinoma. [22, 23,

24, 25]. It has been shown that allogeneic hematopoeitic cell transplant (HCT) can partially rescue the RDEB phenotype. Cells from the bone marrow home to the skin and deposit C7 at the dermal-epidermal junction, greatly improving skin integrity in a subset of patients [26]. However, the molecular mechanism by which this occurs remains unknown. In Chapter 2, to identify sub-populations producing homing signals that could attract bone marrow-derived cells to injured skin, I captured single dermal fibroblasts from patients with severe generalized RDEB and their HLA-matched healthy siblings using the Fluidigm C1 system. In total, 295 patient cells and 248 sibling cells were captured and sequenced. I then used multitask clustering was used to identify novel markers of subpopulations of RDEB fibroblasts.

Treatment for RDEB is limited to palliative care such as bandaging and pain management. One treatment option that addresses the cause of the condition is the use of C7 expressing stem cells or differentiated skin cells to replace C7 at the dermal-epidermal junction and restore the overall integrity of the skin architecture. *Ex vivo* gene therapy for EB has been developed in several pre-clinical studies, with a focus on junctional epidermolysis bullosa (JEB) and RDEB [27]. In 2006, a patient suffering from generalized nonlethal JEB underwent successful transplantation with genetically modified epithelial sheets made from autologous keratinocytes corrected with a classical retroviral vector encoding the b3 chain of laminin 332 [28]. Recently, a similar approach was used in a clinical trial of ex vivo gene therapy for RDEB using the transplantation of autologous epithelial sheets made of primary RDEB keratinocytes genetically modified with a classical retroviral vector expressing the COL7A1 cDNA at Stanford [29]. This type of treatment requires the use of gene therapy (or allogeneic cells), which can be costly and cause adverse reactions in the recipient. In rare cases, genetic reversion causes natural gene correction of the underlying mutation leading to a population or patch of cells that is phenotypically distinct or mosaic. While these cells may be useful clinically, isolating and purifying them has proven difficult. In Chapter 4, I describe a method utilizing synthetic micro RNA (miR) switches, whereby

differences in endogenous miR activity can be exploited to purify mosaic cells in culture. I first show, by targeted sequencing and immunocytochemistry, that C7 is expressed at higher levels in the mosaic compared to the blistered cells. Using miR-sequencing, I identified 10 miRs differentially expressed between mosaic and blistered cells and made miR switches for these. Transfection of the mosaic cells with the miR switches was tested for their ability to separate the corrected from the uncorrected cells. If successful, this technique will be useful for generating pure populations of mosaic cells that can then be used in future clinical applications such as expansion of the naturally-gene corrected cells for autologous hematopoietic cell transplant or skin grafts for areas of the body that fail to heal properly.

C7 is produced by epidermal keratinocytes and dermal fibroblasts. To date, only keratinocytes have been shown to be responsible for revertant mosaicism in RDEB, however, immunostaining data from our lab shows C7 expression by both keratinocytes and fibroblasts in one particular mosaic RDEB patient. This particular patient is a compound heterozygote with a point mutation causing a premature termination codon (PTC) on one *COL7A1* allele and a 16 bp deletion on the other allele predicted to cause a frameshift mutation and downstream PTC. Previous work has indicated that instead of a PTC, this 16 bp deletion causes in-frame exon skipping producing a truncated C7 transcript resulting in a milder dominant dystrophic epidermolysis bullosa (DDEB) phenotype. Interestingly exon-skipping strategies have recently shown promise as therapy for RDEB. This approach relies on the delivery of small antisense molecules to modulate the splicing of the targeted pre-messenger RNA and exclude mutated exons [14]. The first attempt to develop therapeutic exon skipping for RDEB targeted exon 70 of COL7A1, which carries a frequent mutation in the Japanese population [30]. Recently, the efficacy of antisense molecules in skipping exons 73 and 80 was demonstrated, which are frequently mutated in COL7A1, to restore type VII collagen expression and anchoring fibril formation in an in vivo Murine model for RDEB [31]. These small antisense molecules could be delivered subcutaneously

or intravenously to target multiple skin wounds and mucosal lesions in a subset of RDEB patients [14]. In Chapter 5, to determine which cell type is producing enough C7 to restore functionality,potentially due to in-frame exon skipping, I used PacBio sequencing to query the full length of the C7 transcript from fibroblasts and keratinocytes taken from mosaic and blistered sites as well as a wild-type control.

1.2.2 Childhood cerebral adrenoleukodystrophy

The molecular mechanisms responsible for the onset and progression of childhood cerebral adrenoleukodystrophy (ccALD) remain poorly understood. ccALD is one form of X-linked adrenoleukodystrophy (X-ALD), an inherited metabolic storage disorder affecting 1 in 17,000 individuals [32]. X-ALD is caused by mutations in the ABCD1 gene which codes for the ABCD1 protein [33]. ABCD1 is a peroxisomal transporter protein responsible for transporting very long-chain fatty acids (VLCFAs) from the cytosol into the peroxisome for subsequent beta-oxidation [34, 35]. Mutation type and location are not predictive of phenotype, as the same ABCD1 mutation can lead to clinically distinct phenotypes [36, 37, 38, 39, 40] . A more frequent and less severe phenotype, adrenomyeloneuropathy (AMN), presents with demyelination in the long tracts of the spinal cord and progressive axonopathy, usually around the third or fourth decade of life. Heterozygous females will develop similar symptoms by age 60 [41, 42, 43]. ccALD, the most rapidly progressing phenotype, occurs in boys ages 2-12 and is characterized by sudden inflammatory demyelination in the brain and death within a few years [44, 45]. ccALD affects about 40% of males with an ABCD1 mutation [46, 47]. MRI observation of gadolinium enhancement in the brain remains the only method to detect this progression [48, 49, 50, 51, 52]. Infections or head trauma have been described as initiators of the conversion from AMN to ccALD, but typically no extrinsic factor can be identified [53, 54, 55]. Current treatment for ccALD includes hematopoietic cell transplant (HCT), but this must be performed at the

earliest stages of the disease [56, 41, 45, 57].

Much attention has focused on VLCFAs in the search for alternative treatments. While the accumulation of VLCFAs appears to directly contribute to symptoms of AMN, how VLCFAs contribute to the onset or progression of ccALD is unclear [58, 59]. VLCFAs accumulate in many tissue types in X-ALD patients, but this accumulation is not predictive of clinical phenotype [60, 61]. Furthermore, dietary regimens or treatments aimed at reducing the accumulation of VLCFAs (e.g. "Lorenzo's oil") cannot prevent ccALD onset [34, 62, 63], just as immunosuppression cannot prevent the cerebral inflammation seen during ccALD progression [64, 37]. Other biomarkers have been investigated for their potential correlation with ccALD conversion including mitochondrial defects, AMP-activated protein kinases, reactive oxygen species (ROS), and oxidative stress [65, 46, 66, 67, 68, 69]. Antioxidant activity levels of superoxide dismutase in blood plasma have been found to decrease prior to and during cerebral diagnosis [57]. Treatment with the antioxidant N-acetyl-L-cysteine improves survival of patients with advanced ccALD undergoing HCT [70], and oxidative stress levels decrease in patients after HCT [71]. A clinical trial testing a cocktail of antioxidants on patients with AMN has recently been completed though the results have yet to be published [72]. Identification at the molecular level of defects underlying the rapid BBB breakdown seen in ccALD would enable the development of strategies aimed at preventing the onset and progression of ccALD.

The initial blood-brain barrier (BBB) breakdown is thought to be mediated by immune cells (specifically T-cells and to some extent B-cells) translocating from the blood into the brain [73, 74]. Until recently, however, little attention has been paid to the brain endothelium constituting the BBB [75]. X-ALD lacks a suitable mouse model to study the BBB, as mice lacking ABCD1 only develop symptoms of AMN [76]. A human model of the BBB is difficult to obtain, as primary cells isolated from human brain biopsies are not readily available and tend to dedifferentiate upon removal from the in vivo microenvironment [77]. Additionally, immortalized BMEC cell lines display poor barrier properties [78, 79]. To address

these challenges, a system that enables modeling of the BBB through directed differentiation of human induced pluripotent stem cells (hiPSCs) into induced brain microvascular endothelial cells (iBMECs) was recently developed [80, 81, 82, 83]. iBMECs from this system are readily renewable and have been shown to recapitulate important BMEC properties such as junctional protein expression, formation of a tight barrier with physiologically relevant trans-endothelial electrical resistance (TEER) ($\sim$5,000 $\Omega \cdot cm^2$), and multidrug resistance protein efflux activity [80]. This system has been used to model the BBB of other neurological diseases such as Huntington's [84] and to model bacterial interaction with the BBB [85]. In Chapter 3, I used this system to study the BBB of ccALD patients and to ask whether there are differences in barrier function compared to WT controls. The experimental design is outlined in Figure 1.7.

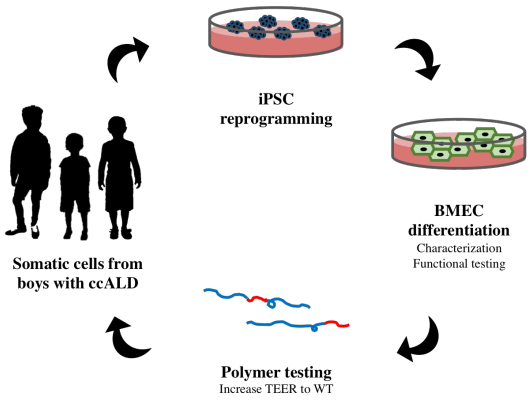


Figure 1.7: **Experimental design.** Somatic cells from clinically confirmed cases of ccALD were reprogrammed into iPSCs before being directed to differentiate into brain microvascular endothelial cells (iBMECs). Diblock copolymers were tested for their ability to improve barrier function in the ccALD-iBMECs.

I used this same system to investigate potential therapeutic interventions to improve defects in barrier function. PEO-PPO-PEO triblock copolymers, called poloxamers or Pluronics, are widely used in biomedical applications due to their biocompatibility and amphiphilicity [86, 87]. Poloxamer 188 (P188; number average molar mass = 8.4 kg/mol and 80 wt% PEO) is approved for human use

in certain applications [88] and has been demonstrated to provide cell membrane stabilization for a panoply of cell and tissue types under various stresses. P188 has been shown to be effective in ameliorating the effects of electropermeabilized skeletal muscle cells in rats [89], skeletal muscle cell necrosis [90, 91], dystrophic heart failure in mice [92], mechanical stress of dystrophic skeletal muscle in mice [93], damaged neuron-like cells in vitro [94, 95], injured primary neurons [96], and acute injury to the BBB in vivo [97, 98, 99]. Moreover, it was recently found that a diblock analog of P188, a PEO-PPO diblock copolymer with one-half the composition and size of P188, can also protect model lipid membranes [100]. PEO-PPO diblock copolymers offer an opportunity to tune the end group on the PPO block, which recent work suggests to be an important molecular parameter in the ability of a PEO-PPO copolymer to confer protection [101]. A systematic in vitro screening of PEO-PPO diblock copolymers identified the polymer $E_{182}P_{16}\underline{t}$ (number average molar mass = 9 kg/mol and 90 wt% PEO; numerical subscripts indicate number of repeat units), which has a hydrophobic tert-butyl (t) end group on the PPO block, to be the most efficacious in stabilizing myoblasts under hypo-osmotic stress and isotonic recovery [85]. Thus, in addition to the commonly used Poloxamer 188, the ability of $E_{182}P_{16}\underline{t}$ to improve iBMEC function was also tested.

1.3 Novel therapies for rare disease

NGS and other technological advances have led to many recent breakthroughs in the understanding of the underlying defects, biological pathways involved, and therapeutic targets of rare diseases, all of which has led to new therapeutic interventions [14]. These discoveries pave the way for improved targeted personalized medicine for rare and frequent diseases. It is likely that a growing number of orphan diseases will benefit from combining new therapies in a near future [14].

1.3.1 Pharmacological and cell-based therapies

It has been a long-standing dream of geneticists to be able to offer afflicted individuals and their families more than counseling and family planning [1]. This is becoming a reality now that we have identified many of the genes responsible for rare diseases as an ongoing goal of translational and regenerative medicine is deciphering disease pathophysiology in order to design novel and specific treatments. Elucidation of disease mechanisms and therapeutic targets allows researchers to design gene, cell, or protein therapies and to use small molecules to specifically oppose disease processes [14]. Human genomics studies continue to guide drug discovery by identifying therapeutic targets while pharmacogenetics leads to improved efficacy and safety of new treatments using information about patient genotypes [14]. Improved understanding of the pathogenesis of skin disorders has led to the development of specific drugs or re-purposing of existing medicines [14].

New pharmacological approaches have been successfully developed in rare Mendelian diseases and in frequent polygenic diseases. They have relied on improved understanding of the pathogenesis of these diseases and the development of specific drugs to reverse the disease pathway [14]. One such approach followed the observation of constitutive activation of IL-1b signaling in cultured keratinocytes from patients with severed generalized EB simplex [19]. Treatment of five children with this disease using topical diacerein, a re-purposed IL-1b inhibitor derived from rhubarb root, significantly decreased blistering in these children for 2 months [19]. Another pharmacological treatment that came about from studying variations on the same disease is the use of losartan, a transforming growth factor-b inhibitor, to treat the inflammation and fibrosis in RDEB [102] and the promotion of wound healing through the recruitment of bone marrow-derived MSCs by recombinant HMGB1 protein [103].

For RDEB patients in particular, a combination of cell, gene, and pharmacological therapies will be required to treat skin and mucosal fragility and their systemic consequences. These therapies could associate intradermal or intravenous infusions of autologous gene-corrected fibroblasts or mesenchymal stromal

cells (MSCs) (potentially derived from iPSCs), grafting of gene-corrected skin equivalents made from keratinocytes, fibroblasts, MSC- or iPSC-derived cells, and molecules reducing skin inflammation and fibrosis like transforming growth factor-b inhibitors [14].

The growing wealth of human genetic data for selecting the best targets should allow us to enter a new age for the treatment of rare diseases with safer and more targeted medicines. Some of these approaches have already been successfully applied to frequent or rare severe genetic skin diseases. It is anticipated that a growing number of patients will benefit from this progress and that new targeted treatments that are currently in pre-clinical development will be translated to clinical medicine in the near future [14].

1.3.2 Gene editing and gene replacement therapies

While interventions such as these and others including enzyme replacement therapy, diet and supplementation with gene products such as clotting factors have been developed for some disorders, a more universal approach for gene replacement or correction would be useful. Especially for rare diseases in which the causative gene is known but the details of the gene function are not easily understood, genetic therapies (either gene replacement or gene repair) can circumvent the need to understand in detail the function of the gene [1]. However, safety is a major concern after the death of Jessie Gelsinger during a gene therapy experiment in 1999 and after four children developed leukemia during a gene therapy (gammaretrovirus) clinical trial for X-linked severe combined immunodeficiency (X-SCID) in 2005 [104, 105, 106]. Improved methods based on non-integrating adeno-associated viral vectors that elicit minimal immune response have allowed gene replacement trials to resume [107, 1]. Very recent successes in hematopoietic stem cells for adrenoleukodystrophy [108] and sickle cell disease [109], liver for hemophilia A [110] and hemophilia B [111], retina for Leber's congenital amaurosis (LCA) [112], skin stem cells for epidermolysis bullosa [113] and systemically

for spinal muscular atrophy [114] suggest a more certain future for viral-mediated gene therapy approaches [1].

Gene addition strategies have shown efficacy in junctional EB and in recessive dystrophic EB (RDEB). TALENs and Cripsr/Cas9 have emerged as highly efficient new tools to edit genomic sequences to create new models and to correct or disrupt mutated genes to treat human diseases [14]. Significant achievements in the area of gene editing for genetic skin disorders have been obtained for junctional, and recessive (RDEB), and dominant dystrophic epidermolysis bullosa over the last 5 years [22]. The first, nuclease-mediated gene editing for COL7A1, was achieved in RDEB patient-derived fibroblasts using TALEN-mediated HR with minimal off-target activity [115].

Induced pluripotent stem cells (iPSCs) represent another therapeutic avenue for the treatment of rare disease, especially in combination with gene editing and knowledge of the key cell types involved. iPSCs not only allow for the generation of patient-specific cellular models of cell types that cannot be obtained by biopsy or are post-mitotic. Skin is an ideal source of tissue for the generation of iPSCs because it is easily accessible with well-established culture conditions for primary fibroblasts and keratinocytes [116]. Their stemness and high proliferative capacity allow easier genetic modification compared with primary cells [14]. iPSCs have been previously derived from RDEB patients [117].

Chapter 2

A Multitask Clustering Approach for Single-cell RNA-seq Analysis in Recessive Dystrophic Epidermolysis Bullosa

2.1 Hypothesis

In recent years, single-cell RNA sequencing (scRNA-seq) has emerged as the dominant method for quantifying transcriptome-wide mRNA expression in individual cells. Single-cell RNA sequencing (scRNA-seq) has been widely applied to discover new cell types by detecting sub-populations in a heterogeneous group of cells. While traditional bulk RNA-seq ignores the differences between individual cells and treats the population of cells as homogeneous, scRNA-seq identifies sub-populations of single cells and can be useful for characterizing sub-population structure, mechanisms of transcription regulation, and understanding disease progression [118] and immunology [119]. A typical scRNA-seq protocol consists of several steps: isolation of single cells and RNA, reverse transcription, amplification, library generation, and sequencing. In addition to the noise and bias that

exist in bulk RNA-seq experiments, issues unique to scRNA-seq include those from biological sources, such as cell-cycle stage or cell size, as well as from technical/systematic sources, such as capture inefficiency, material degradation, sample contamination, amplification biases, GC content, and sequencing depth. These experimental biases and limitations cause uneven coverage of the entire transcriptome and result in an abundance of zero-coverage regions [120, 121]. Since scRNA-seq experiments have lower read coverage/tag counts and introduce more technical biases compared to bulk RNA-seq experiments, the limited number of sampled cells combined with the experimental biases and other dataset specific variations presents a challenge to cross-dataset analysis and discovery of relevant biological variations across multiple cell populations.

Typically, the cost of scRNA-Seq is much higher than bulk RNA-Seq per sample, and thus, scRNA-Seq of a large patient cohort is still prohibitively expensive. When a large number of single-cells from multiple samples are sequenced, more complex batch effects are introduced. Additionally, poorly sampled cell populations could be represented by very few cells for the analysis. To address these challenges, proper integration of multiple scRNA-Seq datasets generated from different experiments is key. When multiple single-cell populations from biological replicates or related samples such as a patient cohort are analyzed to discover the common and sample-specific cell types, technical biases and irrelevant biological variance among independent samples cannot be easily identified and removed from the signal before clustering the single cells. For example, when the scRNA-seq profiles from multiple patients are pooled together for clustering, the clusters will highly overlap with the division of the single cells by the sample origins rather than similar types such as pathogenic cells vs normal cells.

I hypothesized that applying multitask learning to scRNA-seq data would improve clustering and marker identification. A multitask learning method with embedded feature selection was introduced to simultaneously capture the differentially expressed genes among cell clusters and across all cell populations to achieve better single-cell clustering. The key advantage of multitask clustering

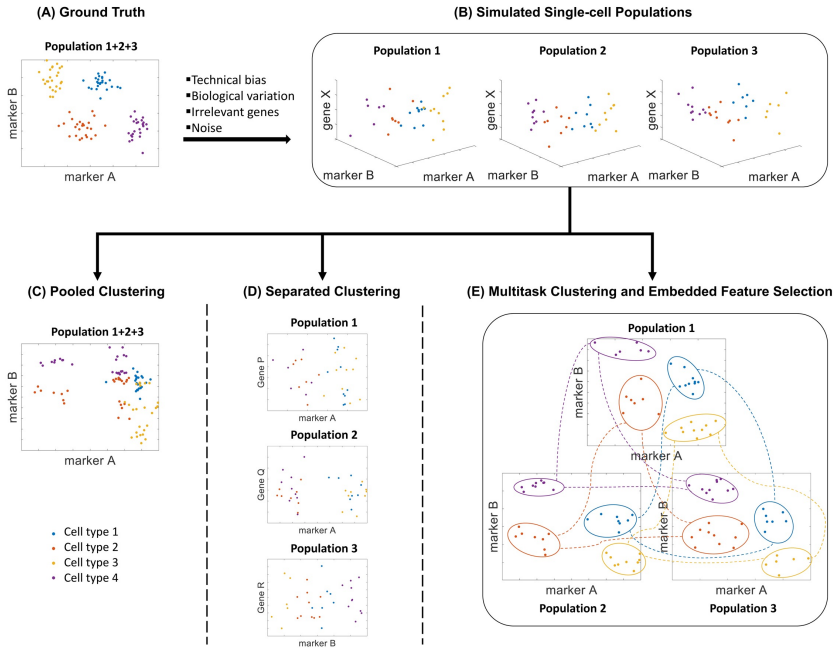


Figure 2.1: **Strategies for clustering multiple single-cell populations**. In the example, four cell types are shown using four colors. **(A)** Ground Truth. 2D plot of a pool of single cells combined from 3 single cell populations of identical distributions separated by true marker genes A and gene B. **(B)** Simulated Single-cell Populations. 3D plots of the three single-cell populations separated by marker genes A, B and non-marker gene X. The simulation data are generated from the ground truth data with rotation and scaling to represent technical biases and biological variation with 998 random genes in addition to gene A and gene B (1000 genes in total). Additional noise is also introduced. Three different clustering strategies are shown below in **(C)**, **(D)** and **(E)**. **(C)** Pooled Clustering. The 2D plot with true marker genes A and B on pooled data that simply combines 3 single-cell populations together for clustering is shown. Even with the correct marker selection, cells from different types are still mixed because of the rotation, scaling and noise. **(D)** Separated Clustering. The 2D plot on each individual cell population is shown. With the limited single-cell sample size and skewed cell-type distribution, incorrect marker genes may be selected, shown as genes P, Q and R. **(E)** Multitask Clustering and Embedded Feature Selection. The proposed multitask clustering can identify both true marker genes and correctly cluster the individual cells into their respective types in each population. The clustering of each dataset reinforces the results of the other two datasets shown as the connected clusters across the three experiments.

is the use of multiple single-cell populations to leverage the sample size limitation in each individual dataset while allowing dataset-specific variations among the same cell types across the datasets. To illustrate the objective, Figure 2.1 shows a simulation example of scRNA-seq data of 100 single cells from three cell populations (n = 33, 33 and 34) with 1000 expressed genes. Among the 1000 genes, gene A and gene B are the hidden markers that are differentially expressed across the four cell types (indicated by four different colors). In the ideal scenario, there is no technical bias and the marker genes are known as shown in the ground truth in Figure 2.1(A). Figure 2.1(B) shows the single-cell datasets after biological variation, technical biases, and noise are introduced.

The data distributions are very different across the three cell populations after the rotation, re-scaling and addition of noise. It is also challenging to identify the true marker genes with a limited number of samples in each population. Simply pooling the single-cell data from the three populations together will confuse the clustering, even with the correct marker genes identified (Figure 2.1(C)). Conversely, separated clustering on each single-cell population suffers more from the biological variation as the number of single cells is not sufficient in each individual analysis to identify the true maker genes (Figure 2.1(D)). As shown in Figure 2.1(E), variance-driven multitask clustering of single-cell RNA-seq data (scVDMC) utilizes expression patterns of different single-cell populations with shared cell-type markers and corresponding similar clusters for better integration.

2.2 Materials and methods

In this section, the model and the algorithm of variance-driven multitask clustering of single cells (scVDMC) is first introduced and then the parameter selection for scVDMC and related work in scRNA-seq clustering is discussed. Here also the methods for the generation of the in-house Recessive Dystrophic Epidermolysis Bullosa (RDEB) scRNA-seq dataset and the flow cytometry experiments are described.

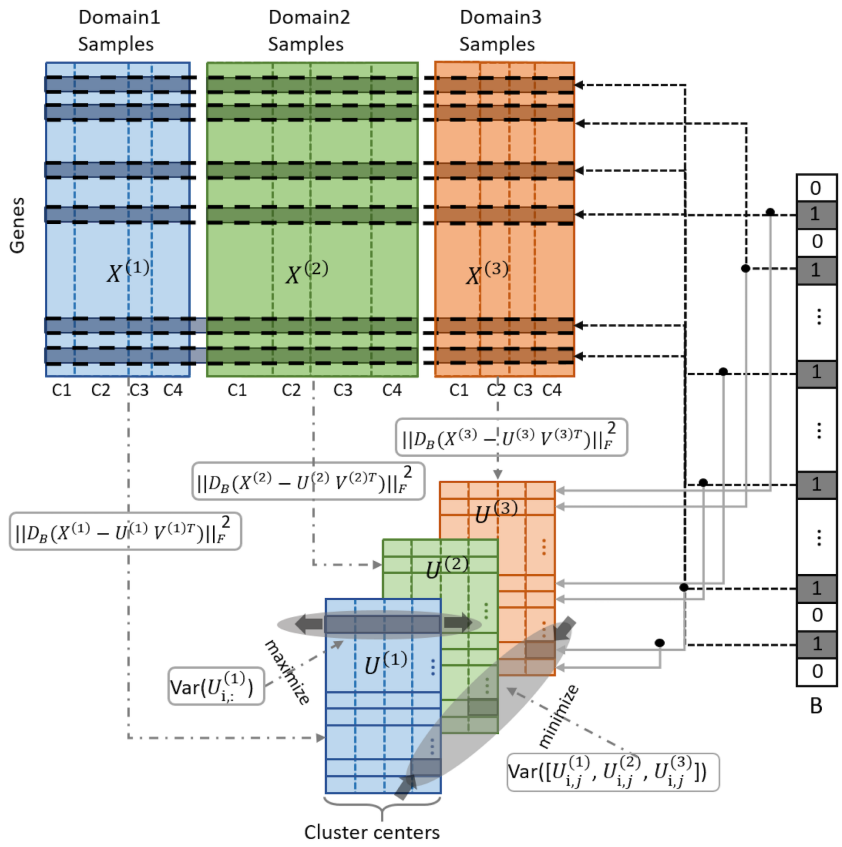


Figure 2.2: **Variance-driven multitask clustering model**. Three domains (single-cell populations) are clustered into four cell clusters (C1-C4) in multitask clustering. The samples in each domain are in four clusters separated by the vertical bars. Each dataset is clustered by factorization of the data matrix by the selected genes (with indicator 1 in B) common to the three domains. Two types of variance are controlled, 1) the variance among the cluster centers in the same domain are maximized for better cluster separation shown as a shadowed row; and 2) the variance among the shadowed cluster centers across the domains are minimized to match the similar clusters across the domains.

2.2.1 A multitask clustering and feature selection model

Assume a total of D domains with each domain representing a single-cell population for clustering. Let matrix $X^{(d)} \in \mathbb{R}^{m \times n^{(d)}}$ denote RNA-seq gene expression values from domain $d \in \{1, 2, \ldots, D\}$, where m is the number of features (genes)

and $n^{(d)}$ is the single-cell sample size of domain d. Let $U^{(d)} \in \mathbb{R}^{m \times k}$ denote the cell-type cluster centers, vector $Y^{i,j} = [U^{(1)}_{i,j}, U^{(2)}_{i,j}, \ldots, U^{(D)}_{i,j}]^T$ stack the (i,j)-th entry of every $U^{(d)}$ and the binary matrix $V^{(d)} \in \{0,1\}^{n^{(d)} \times k}$ denote the assignments of each single-cell to the clusters, where k is the number of cell types (clusters). With the binary vector $B \in \{0,1\}^m$ denoting the indicators of feature selection (1: selected and 0: not selected) and D_B denoting the diagonal matrix with B on the diagonal, scVDMC model outlined in Figure 2.2 is defined as:

$$\begin{aligned} \underset{U^{(d)},V^{(d)},B}{\text{minimize}} \quad & \frac{1}{2}\sum_{d=1}^{D} ||D_B(X^{(d)} - U^{(d)}{V^{(d)}}^T)||_F^2 - w\sum_{d=1}^{D} B^T \text{Var}(U^{(d)}) + \alpha \sum_{i,j} B_i \text{Var}(Y^{(i,j)}) \\ \text{subject to} \quad & \sum B = \lambda, \\ & \sum_j V^{(d)}_{i,j} = 1, \forall i = 1, 2, \ldots, n^{(d)}, \ \forall d = 1, 2, \ldots, D, \end{aligned} \tag{2.1}$$

where w and $\alpha > 0$ are hyper-parameters to balance the three error terms: the reconstruction error, the cluster center separation in each cell population, and the variance of the cluster centers across the different single-cell populations. $\lambda \in \mathbb{Z}^+$ is the predefined number of features to be selected. $||D_B(X^{(d)} - U^{(d)}{V^{(d)}}^T)||_F^2$ in equation (2.1) denotes the reconstruction error of the classic k-means clustering as matrix factorization with D_B selecting marker genes by B, i.e. the reconstruction error is only measured on the marker genes by ignoring the irrelevant (non-selected) genes. The second term $B^T \text{Var}(U^{(d)})$ is introduced to maximize the separation of the cluster centers, where $\text{Var}(U^{(d)})$ is defined as a vector in which each element is the variance of the vector $U^{(d)}_{i,:} \in \mathbb{R}^{k \times 1}$ [122]. The third term $\text{Var}(Y^{(i,j)})$ denotes the variance of the vector $Y^{(i,j)}$, which is introduced to require similar gene expression centers across different single-cell populations. Note that the reconstruction error encourages selection of low expression genes since the errors are usually smaller on smaller values while the second variance term encourages selection of high expression genes since the variances tend to be larger on larger values. Together as the sum over all the domains, the cost function

provides a balanced error on the compactness and separation of the clusters of the cell types tuned by feature selection across all the domains. The unique but similar cluster centers in each domain preserves the unique expression patterns while the features are selected as common marker genes for different cell types. For the three hyper-parameters in equation (2.1), λ (the number of marker genes) is typically a small number based on prior knowledge of the cell types, and the selection of balancing weight w and α is discussed later in this section.

2.2.2 Alternating updating algorithm

Algorithm 1 scVDMC algorithm

Input: $X^{(d)}, \alpha, k, w, \lambda, d = 1, 2, \ldots, D$
output: $U^{(d)}, V^{(d)}, B$
Initialize $U^{(d)}$ and $V^{(d)}$.
repeat
    compute B with integer linear programming in equation (2.7)
    for $d = 1, 2, \ldots, D$ **do**
        solve $V^{(d)}$ by equation (2.2)
        solve $U^{(d)}$ by (2.6)
    end for
until $U^{(d)}$, $V^{(d)}$ and B converge
return $U^{(d)}$, $V^{(d)}$ and B

The full scVDMC algorithm is shown in Algorithm 1. The goal is to minimize the cost function in equation (2.1) to obtain the optimal $U^{(d)}$, $V^{(d)}$ and B. An alternating update strategy is employed to solve the optimization problem. First, the feature selection B, all the cluster centers $U^{(i)}, i = 1, 2, \ldots, D$ and all other $V^{(i)}, i \neq d$ are fixed to obtain a certain $V^{(d)}$.

$$\begin{aligned} \underset{V^{(d)}}{\text{minimize}} \quad & \frac{1}{2}||D_B(X^{(d)} - U^{(d)}{V^{(d)}}^T)||_F^2 \\ \text{subject to} \quad & \sum_j V_{i,j}^{(d)} = 1, \quad \forall i = 1, 2, \ldots, n^{(d)}. \end{aligned} \tag{2.2}$$

This is equivalent to assigning samples to the nearest centers $U^{(d)}$ by the Euclidean distance in the features selected by B, where each column of $D_B X^{(d)}$ is a sample and each column of $D_B U^{(d)}$ is a center. Then the distance of a sample to every center is calculated and the nearest center is chosen to assign 1 to the corresponding $V^{(d)}$. The time complexity for assigning each sample to one of the k clusters over the λ marker genes will be $O(n \times k \times \lambda)$.

Next, the feature selection B is fixed in all clustering assignments $V^{(i)}, i = 1, 2, \ldots, D$, and all other $U^{(i)}, i \neq d$, to solve a certain $U^{(d)}$, rewritten as:

$$\underset{U^{(d)}}{\text{minimize}} \ \frac{1}{2}\sum_{i=1}^{m} B_i||(X_{i,:}^{(d)} - U_{i,:}^{(d)}V^{(d)^T})||_2^2 - w\sum_{i=1}^{m} B_i \text{Var}(U_{i,:}^{(d)}) + \alpha \sum_{i,j} B_i \text{Var}(Y^{(i,j)}), \tag{2.3}$$

where $\text{Var}(U_{i,:}^{(d)})$ is the variance of vector $U_{i,:}^{(d)}$ defined as

$$\begin{aligned} \text{Var}(U_{i,:}^{(d)}) &= \frac{1}{k}(U_{i,:}^{(d)} - \frac{U_{i,:}^{(d)}\mathbf{1}_k\mathbf{1}_k^T}{k})(U_{i,:}^{(d)} - \frac{U_{i,:}^{(d)}\mathbf{1}_k\mathbf{1}_k^T}{k})^T \\ &= \frac{1}{k}U_{i,:}^{(d)}(\mathbf{I}_k - \frac{\mathbf{1}_k\mathbf{1}_k^T}{k})(\mathbf{I}_k - \frac{\mathbf{1}_k\mathbf{1}_k^T}{k})^T U_{i,:}^{(d)^T} \\ &= \frac{1}{k}U_{i,:}^{(d)}(\mathbf{I}_k - \frac{\mathbf{1}_k\mathbf{1}_k^T}{k})U_{i,:}^{(d)^T}, \end{aligned} \tag{2.4}$$

where $\mathbf{I}_k$ denotes the identity matrix of size k and $\mathbf{1}_k$ is a length k column vector of all ones. Similarly, this becomes

$$\text{Var}(Y^{(i,j)}) = \frac{1}{d}Y^{(i,j)T}(\mathbf{I}_d - \frac{\mathbf{1}_d\mathbf{1}_d^T}{d})Y^{(i,j)}. \tag{2.5}$$

As shown in Appendix S1, the analytical solution of equation (2.3) when $B_i = 1$

is

$$U_{i,:}^{(d)^T} = (V^{(d)^T}V^{(d)} - \frac{2w}{k}\Psi + \frac{2\alpha k\Phi_{d,d}}{d}\mathbf{I}_k)^{-1}(V^{(d)^T}X_{i,:}^{(d)^T} - \frac{2\alpha k}{d}\sum_{l\neq d}\Phi_{dl}U_{i,:}^{(l)T}). \tag{2.6}$$

The time complexity is $O(k^3)$ for the matrix inversion and $O(n\times k^2)$ for computing $V^{(d)^T}V^{(d)}$. Since the matrix inversion is common to all the genes and only needs to be computed once, the total time complexity is only $O(n \times k \times \lambda)$.

Finally, to update binary vector B, all $U^{(d)}$ and $V^{(d)}$ are fixed to optimize

$$\begin{aligned} \underset{B}{\text{minimize}} \quad & \sum_{i=1}^{m} B_i(\frac{1}{2}\sum_{d=1}^{D}||(X_{i,:}^{(d)} - U_{i,:}^{(d)}V^{(d)^T})||_2^2 - w\sum_{d=1}^{D}\text{Var}(U_{i,:}^{(d)}) + \alpha\sum_{j=1}^{k}\text{Var}(Y^{(i,j)})) \\ \text{subject to} \quad & \sum B = \lambda, \end{aligned} \tag{2.7}$$

which is a standard constrained linear binary integer programming problem that can be easily solved by sorting the coefficients of B and taking the top λ entries. The time complexity is $O(m \times n \times k)$ for computing the construction error terms, $O(D \times m \times k)$ for computing the variances and $O(m \log m)$ for sorting the coefficients. The overall time complexity is $O(m \times n \times k)$ assuming $n \times k > \log m$. Thus, the total time complexity of each iteration in Algorithm 1 will be $O((m + \lambda) \times n \times k)$, which is comparable to k-means when $\lambda << m$.

2.2.3 Parameter selection

There are four hyper-parameters to tune for the scVDMC algorithm, α and w: weights of the two variance terms, k: the number of clusters and λ: the number of marker genes. The strategies for tuning α, w and k are described below assuming that λ can be approximately informed by prior knowledge of the cell types.

Tuning α: The role of α is to weight the cost term on the cross-domain variance of the cluster centers. The larger the α the more similar the cluster centers are across

the domains. Ideally, α should be relatively small to allow smaller reconstruction error but yet meet the consistency requirement across the domains. The strategy is to start with a small α and measure the total difference between the cluster centers of the corresponding cluster across the domains, and then increase α to reduce the difference until the total difference does not change much. This selection can also be achieved by visualization of the cluster centers with Principle Component Analysis (PCA) or other dimension reduction methods. After clustering, the data can be projected in each domain into the first two PCs. The distance between the cluster centers of the same cluster in each domain can be compared for choosing an appropriate α. Several examples are shown later in the experiments.

Deriving the upper bound of w: Equation (2.3) is a sum of a few quadratic terms of variable $U_{i,:}^{(d)}$. The global minimum of $U_{i,:}^{(d)}$ can be solved in closed-form if the Hessian below is positive semi-definite,

$$H = V^{(d)^T} V^{(d)} - \frac{2w}{k}\Psi + \frac{2\alpha k \Phi_{d,d}}{d}\mathbf{I}_k. \tag{2.8}$$

In the following, it shows that an upper bound on w will guarantee that H is positive semi-definite. By Gershgorin circle theorem[1], the sufficient condition of $H \succeq 0$ is $H_{ii} - \sum_{j\neq i} |H_{ij}| \geq 0$ for $\forall i$. This is equivalent to stating that H is diagonally dominant and only has non-negative diagonal entries. H can be rewritten as follows,

$$\begin{aligned} H_{ii} &= c_i + \frac{2w(1-k)}{k^2} + \frac{2\alpha k(d-1)}{d^2},\ \forall i = 1, ..., k \\ H_{ij} &= \frac{2w}{k^2},\ \forall i \neq j, \end{aligned}$$

where c_i is the i^{th} diagonal entry of matrix $V^{(d)^T} V^{(d)}$, i.e., the size of cluster i.

[1]For any eigenvalue δ of matrix H, $|\delta - H_{ii}| \leq \sum_{j\neq i} |H_{ij}|$ for $\forall i \Longleftrightarrow H_{ii} - \sum_{j\neq i} |H_{ij}| \leq \delta \leq H_{ii} + \sum_{j\neq i} |H_{ij}|$.

This becomes

$$c_i + \frac{2w(1-k)}{k^2} + \frac{2\alpha k(d-1)}{d^2} \geq \frac{2w(k-1)}{k^2}$$

and thus,

$$w \leq \frac{k^2 c_{min}}{4(k-1)} + \frac{\alpha k^3(d-1)}{2d^2(k-1)}$$

where c_{min} is the minimum of c_i, $\forall i = 1, ..., k$. Since $c_{min} \geq 1$ (no empty cluster), a loose upper bound of $w = \frac{k^2}{4(k-1)} + \frac{\alpha k^3(d-1)}{2d^2(k-1)}$ is obtained. In all the experiments, w is set to be smaller than the upper bound for feasible implementation.

Determining the number of clusters k: The number of clusters k is selected by an "elbow" plot of the within-clusters sum of squares T_s computed as follows:

$$T_s = \sum_{d=1}^{D} ||D_B(X_{i,:}^{(d)} - U_{i,:}^{(d)} V^{(d)^T})||_2^2. \tag{2.9}$$

T_s represents the amount of variance to minimize for better clustering. Larger k will lead to smaller T_s. By plotting T_s under different options of k, the best k at the so-called "elbow" of the curve can be selected. Supplementary Figures S6 and S7 show the "elbow" plot on two datasets in the experiments. In addition, when an empty cluster is created, the calculation of cluster center variance will be invalid. To address the possible issue, simple splitting procedure is used to handle empty clusters. Specifically, if there is an empty cluster in $V^{(d)}$ (i.e. the whole column is 0), the largest cluster is randomly split into two clusters. This procedure is repeated until there are exactly k clusters. This strategy is similar to commonly used k-mean rerun when a cluster center is collapsed on a single data point or no data point.

2.2.4 scRNA-seq of RDEB cohort

To identify sub-populations producing homing signals that could attract bone marrow-derived cells to injured skin, single dermal fibroblasts from six patients with severe generalized RDEB and their HLA-matched healthy siblings were captured using the Fluidigm C1 system. The demographics information of the patients and donors are shown in Supplementary Table S1.

Cell culture: Dermal fibroblasts from patients with severe generalized RDEB and their human leukocyte antigen (HLA) matched healthy siblings were obtained from skin biopsies and cultured in DMEM high glucose (Thermo Fisher Scientific) containing 10% fetal bovine serum (MilliporeSigma), 1% Pen/Strep (Thermo Fisher Scientific), 1% L-glutamine (Thermo Fisher Scientific), and 1% MEM NEAA (Thermo Fisher Scientific). For sub-culture, the medium was removed and cells were washed with 1X PBS (Thermo Fisher Scientific) and detached using Trypsin/EDTA (Thermo Fisher Scientific). Experiments were performed with fibroblasts at passages 4-9.

Single-cell capture and RNA-seq: Fibroblasts were collected by trypsinization and resuspended in 5 μL of fibroblast medium for loading into the capture chip. The medium- (10-17 μm diameter) and large-size (17-25 μm diameter) chips were used to capture cells with the C1 system (Fluidigm). Cells were loaded at a concentration of 2.5 x 10^5 per μL and stained with the Live/Dead Viability/Cytotoxicity kit (Thermo Fisher Scientific). Cells were imaged with phase-contrast and fluorescence microscopy to assess cell number and viability at each capture point. Capture sites with single, live cells were selected while capture sites with multiple, no, or an unclear number of cells were excluded from further analysis. Images for each single-cell used in this study are available upon request. In total, 295 patient cells and 248 sibling cells were selected. On the device, cDNA was created from the selected cells using the SMARTer Ultra Low RNA kit designed for the C1 system (Clontech). mRNA libraries were constructed using the Nextera XT kit (Illumina) according to the manufacturer's protocol. The libraries were sequenced on an Illumina MiSeqv3 with 75bp paired-end reads to a depth of

19-22 million reads per lane. Target sequencing depth for each library was 200K reads.

RDEB-WT pairs	RDEB cells	Avg. reads	WT cells	Avg. reads
1	41	248,929	20	205,216
2	72	200,961	46	241,966
3	54	138,598	37	146,610
4	36	83,513	51	86,483
5	46	181,263	47	176,929
6	46	175,346	47	170,307

Table 2.1: For each RDEB or WT individual, the number of single-cells used for downstream analysis is indicated as well as the average number of reads for the single-cells from each individual.

Processing of RNA-seq data: Paired-end 75bp reads were mapped to the UCSC human transcriptome (hg19) using Bowtie2 (version 2.2.4) and Tophat (version 2.0.9). Gene expression levels were calculated using Cuffquant (Cufflinks version 2.2.1 with parameters -u -max-bundle-frags 10000000) and Cuffnorm (Cufflinks version 2.2.1). FPKM values as estimated by Cufflinks were added a value of 1 (to avoid zeros) and log2 transformed. Nine single-cell samples with low read counts (< 50K) were removed and two single-cell samples sequenced as population controls with high read counts (> 1.5M) were sub-sampled (random sub-sampling, 10% of total reads). 11 single-cell samples were excluded as outliers. Lowly expressed genes (average log2 (FPKM) < 1.5) were excluded from further analysis. The remaining 543 single-cell samples met the requirement of expressing at least 2,000 of these remaining 5,196 genes. For each individual, the number of single-cells used in the analysis and the average number of reads for those single-cells is summarized in Table 2.1. The total number of the reads and the number aligned reads in each cell are also shown (Supplementary Figure S3).

Flow cytometry: Fibroblasts were collected by trypsinization (as above) and resuspended in fibroblast medium. A BD Cytofix/Cytoperm^{TM} kit (BD Biosciences) was used to prepare the cells for intracellular staining. Cells were fixed for 15 min with 150 μl Fixation/Permeabilization solution before being resuspended in 300 μl

1X BD Perm/Wash Buffer and incubated at 4°C for 20 min. Primary antibodies (Supplementary Table S2) were diluted in 100 μl 1X BD Perm/Wash Buffer and cells were resuspended in this for 20-30 min at 4°C, followed by one wash with 500 μl 1X BD Perm/Wash Buffer. Secondary antibodies (Supplementary Table S3) were diluted in 300 μl 1X BD Perm/Wash Buffer and cells were resuspended in this for 20-30 min at 4°C, followed by one wash with 500 μl 1X BD Perm/Wash Buffer, and resuspension in 300 μl 1X BD Perm/Wash Buffer. Flow cytometry experiments were carried out on a BD LSRII system equipped with FACsDiva 8.0 software (BD Biosciences) and analyzed using FlowJo (Tree Star Inc.).

2.2.5 Related work

Most existing methods focus only on sub-population clustering and differential gene expression detection among the learned cell clusters with one (pooled) cell population. Some of these methods were directly adopted from traditional bulk RNA-seq analysis and/or classical dimension reduction algorithms such as Principal Component Analysis [123, 124, 125], hierarchical clustering [126], t-SNE [127, 128, 129], Independent Component Analysis [130] and Multi-dimensional Scaling [131]. Other methods focus on special properties of scRNA-seq data, such as high variance and uneven expressions. For example, SNN-Cliq [132] uses a ranking measurement to get reliable results on high dimensional data; [133] proposed a special dimension reduction method to handle the large amount of zeros in scRNA-seq; [134] proposed a Latent Dirichlet Allocation model with latent gene groups to measure cell-to-cell distance; CellTree method [134] clusters single cells by a detected tree structure outlining the hierarchical relationship between single-cell samples to introduce biological prior knowledge; Seurat [135] was proposed to infer cellular localization by integrating single-cell RNA-seq data with *in situ* RNA patterns; and more recently a consensus clustering approach SC3 [136] was proposed to improve the robustness of clustering through combining multiple

clustering solutions by consensus.

Mixed multiple batch strategies [126, 137] have been proposed to reduce the technical variance, which does not directly improve clustering. However, multitask clustering with an embedded feature selection has not been previously applied to scRNA-seq data analysis.

Datasets	# of cells	# of clusters	# of domains	# of cells in each domain
mESC	250	3	3	81:90:79
Lung	77	4	3	20:34:23
PBMC	27,302	10	3	10000:7783:9519
RDEB	543	4	6	61:118:91:87:93:93

Table 2.2: Four datasets used in the experiments.

In the experiments, scVDMC was applied to two small scRNA-seq datasets: mouse embryonic stem cell (mESC) data [138] and mouse embryonic lung epithelial cell (Lung) data [139], and one large-scale Droplet-based scRNA-seq peripheral blood mononuclear cells (PBMC) data [140] . scVDMC was also applied to an in-house Recessive Dystrophic Epidermolysis Bullosa (RDEB) data to detect RDEB relevant cell types and marker genes. The statistics of the four datasets are shown in Table 2.2.

2.2.6 Experimental design

scVDMC was compared with six baseline methods: (1) k-means clustering on each domain separately, (2) pooling all domains and applying k-means clustering, (3) SNN-Cliq [132], (4) CellTree [134], (5) Seurat [135] and (6) SC3 [136]. Pooled k-means (2) was used to obtain the initialization for scVDMC.

To apply the SNN-Cliq method [132], the provided MATLAB code was used to transform the data into the SNN graph, then the Python code was used to produce the clustering result by ranking measurement. There are three hyper-parameters: k (size of the nearest neighbor list), r (parameter for quasi-clique finding, range (0,1]), and m (parameter for cluster merging range (0,1]). Multiple combinations

of the three hyper-parameters were tested using $k = 3, 5, 7$, $r = 0.1, 0.2, \ldots, 0.9$ and $m = 0.1, 0.2, \ldots, 0.9$. The program was set to require annotation of all the data instead of leaving singletons unlabeled ($-n$). Since SNN-Cliq identifies the number of clusters automatically, only the results with the correct number of clusters were reported.In all experiments with SNN-Cliq, genes with low expression were removed and the data was log-transformed, as recommended in [132].

To apply the CellTree method [134], the provided R package was used to first fit a Latent Dirichlet Allocation (LDA) model with the default method (joint MAP estimation) to choose the number of topics followed by learning a pair-wise distance for all cells. Next, hierarchical clustering with four different methods was run for computing cluster distance ('ward', 'complete', 'single', 'average') and the best clustering results were selected.

To apply Seurat [135], Seurat v2.0 R package was downloaded from SATIJA LAB. The scRNA-seq data were converted into the required format (gene index | cell index | gene expression) as the input. The parameter "Resolution" tunes the granularity of the downstream clustering, with increased values resulting a larger number of clusters. A range [0.5,1.5] was tested to get the exact number of clusters for comparison with other methods. The reported result of Seurat is computed with the resolution parameter that gives the exact number of clusters and the lowest error.

To apply the SC3 [136], the SC3 v1.7.2 R package was downloaded from Bioconductor. All parameters in SC3 are set to default. In the experiments with more than 5000 instances for clustering, the SVM mode will be triggered to run a second stage supervised learning to improve the scalability.

To further test separated cluster, pooled clustering and SC3 combined with feature selection, genes with larger variance were chosen as the marker genes. Since the other three baselines use a different strategy for clustering and do not provide marker-gene selection, the clustering result for these three baselines was made the focus. The true cluster labels are obtained as the validated clusters with high confidence in the mESC data [138] and Lung data [139], and the known

PBMC populations from donor A sorted with FACS analysis [140].

2.2.7 Experiment on mouse embryonic stem cell data

The single-cell expression data for 250 mESCs [138] was downloaded from the European Bioinformatics Institute's (EBI) ESpresso database. These 250 mESCs cultured in serum conditions were captured using the Fluidigm C1 on three different days from three different passages (biological replicates, n = 81, 90, and 79). After removing genes expressed uniformly within a single replicate, 12,114 genes remained. To tune α for scVDMC, the positions of the cluster centers across the domains were examined and visualized by PCA in Supplementary Figure S4(A) and (B). Based on the visualization, $\alpha = 0$ and 1 was chosen since the relative positioning of cluster centers was similar in all the three domains. For the SNN-Cliq method, genes with log-transformed average expression less than 20 were removed.

Figure 2.3(A) shows the clustering results. Compared with the six baselines, scVDMC shows a consistently lower error with different choices of λs. Within a reasonable range of λ, such as from 20 to 200, scVDMC shows significant improvement compared with the baseline methods. When λ is too small, such as 10 genes selected, there are not enough markers to capture the difference among the cell types such that the error is larger. When λ is too large, scVDMC will consider almost all the genes and the variance selection will not play a role. As such, scVDMC will eventually degrade into separated k-means and the error will also increase. As shown in Supplementary Figure S1(A), it is worth noting that the results are less sensitive to the choice of the parameter w, for which the upper bound for w is $\frac{9}{8}$ in this case. It is also interesting that the CellTree method performed better than both pooled and separated k-means, while SNN-Cliq and SC3 performed better than separated k-means but worse than pooled k-means. Under various tuning of the parameters, Seurat still performed poorly on this dataset. Both separated k-means and pooled k-means performed much worse

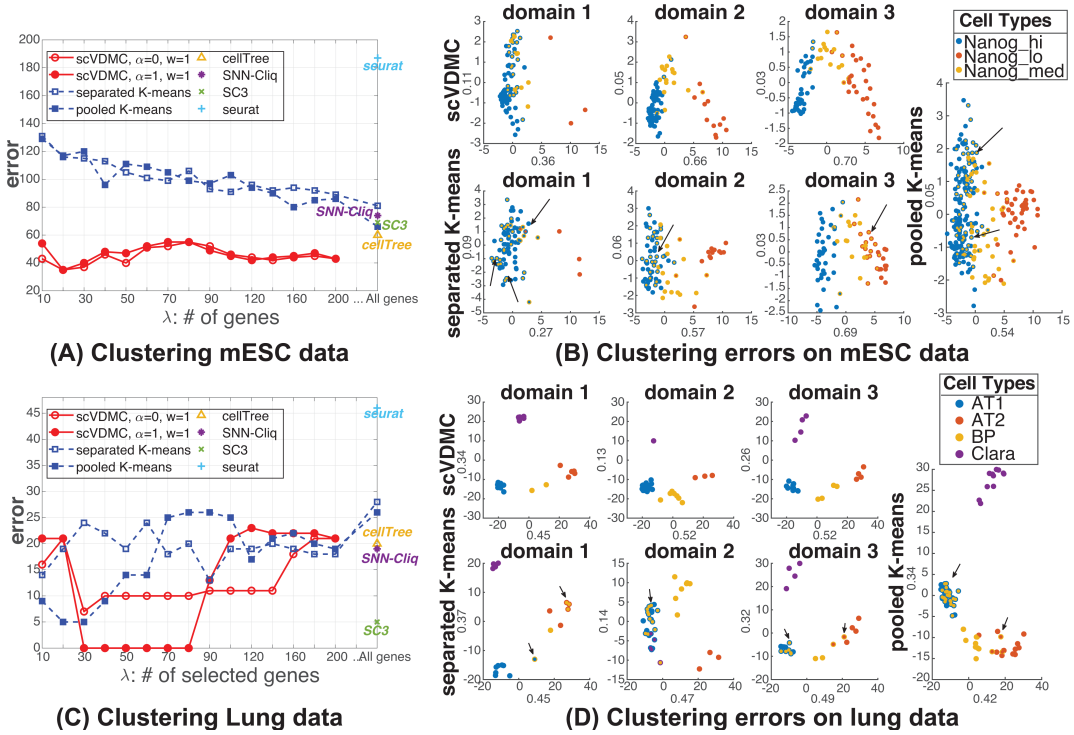


Figure 2.3: **Clustering performance on mESC and Lung datasets**. (A) & (C) show the clustering results of the scVDMC algorithm compared with the baseline methods. Pooled k-means, separated k-means and scVDMC are tested with varying numbers of selected marker genes. Seurat, cellTree, SNN-Cliq and SC3 are tested using all the genes as input to the software/program. (B) & (D) show the PCA of scVDMC, pooled k-means, and separated k-means results on the selected top 20 marker genes. PCA is applied on each individual domain for separated k-means and the combined data for pooled k-means and scVDMC. For each dot, the layer (outer) color indicates the true cell type, while the inner color indicates the predicted cell type. The error is measured on the best one-to-one matching between the detected clusters and the true cluster. The hyper-parameters for scVDMC are $\lambda = 20, w = 0.1, \alpha = 0.5$ on the mESC dataset and $\lambda = 50, w = 0.1, \alpha = 1$ on the Lung dataset.

with the feature selection by variance, indicating that simple feature selection strategies will not identify correct markers in this dataset. Running scVDMC with $\alpha = 1$ performed the best when 20 marker genes are selected but the overall performance is very similarly as running with $\alpha = 0$, indicating that the control of the cross-domain variance could play a role in improving the results. However, since the cluster centers are already not very different when running with $\alpha = 0$, the improvement will only be marginal. Figure 2.3(B) shows the detailed clustering errors by scVDMC, pooled k-means and separated k-means. Compared with the pooled k-means and separated k-means, scVDMC captures relatively high variance in the leading principle components and achieves improved clustering in every domain (fewer mixed-color dots). In Supplementary Figure S2(A), the convergence of scVDMC by the number of iterations is also shown.

Analysis of the mESC transcriptome data using scVDMC yielded comparable results on marker gene selection in the original paper [138] as well as pooled and separated k-means. Both analyses were able to detect and highly rank the known markers of differentiation *Krt8, Krt18, Anxa1, Anxa3*, and *Acta1*. Further, scVDMC detected several additional genes that pooled k-means, separated k-means and the original paper did not. These included *Cav1*, which is required for normal lung development [141] and *Dsp*, variants of which are associated with idiopathic pulmonary fibrosis [142].

2.2.8 Experiment on lung epithelial single-cell data

The single-cell expression data for 80 embryonic mouse lung epithelial cells [139] was downloaded. These 80 single-cell samples were taken from three different mice (biological replicates, n = 20, 34, and 23) and contained five cell types: ciliated, Clara, AT1, and AT2 cells, as well as a bi-potential progenitor (BP). Since only one replicate contained ciliated cells, these were removed from the analysis, leaving 77 single-cell samples. After removing genes expressed uniformly within a single replicate, 7,357 genes remained. To tune α for scVDMC, the positions

of the cluster centers across the domains were examined and the visualization by PCA is shown in Supplementary Figure S4(C) and (D). $\alpha = 1$ is chosen as the optimal parameter to achieve similar relative positioning of cluster centers in all the three domains. For the SNN-Cliq method, genes with log-transformed average expression less than 2 were removed.

With the limited number of single-cell samples in this dataset, scVDMC still much improved clustering over the baselines in the range of $\lambda \in [30, 80]$ shown in Figure 2.3(C). In Figure 2.3(D), PCA plots of the top 30 genes show a trend similar to the ESC dataset, where scVDMC's top genes capture more variance and show less clustering error. Both SNN-Cliq and CellTree performed better than pooled k-means and separated k-means, with SNN-Cliq leading CellTree by a very small margin. Similarly, Seurat also performed poorly while SC3 performed well on the dataset with only 5 mistakes. It is also interesting to observe that running scVDMC with $\alpha = 1$ performed significantly better than running with $\alpha = 0$, indicating that the control of the cross-domain variance played an important role in improving the results. Since the cluster centers are very different when running with $\alpha = 0$, the improvement is significant. Another interesting observation is that the clustering performance is more sensitive to the number of marker genes to select by scVDMC. In particular, selection of 20-80 genes with scVDMC ($\alpha = 1$) will give the optimal clustering results while selection of more than 90 genes will give much higher error. This is due to the small clusters in this dataset (e.g. purple cluster in domain 2 and yellow cluster in domain 1), which could be sensitive to the number of selected genes in low-read-coverage samples. Thus, the error will be more sensitive to the gene selection in this small dataset. On this dataset, both separated k-means and pooled k-means performed better with the feature selection by variance but never achieved zero clustering error as scVDMC does. As shown in Supplementary Figure S1(B) and S2(B), scVDMC behaved similarly by the choices of the w parameters and the convergence.

Analysis of the mouse lung epithelial transcriptome data using scVDMC yielded

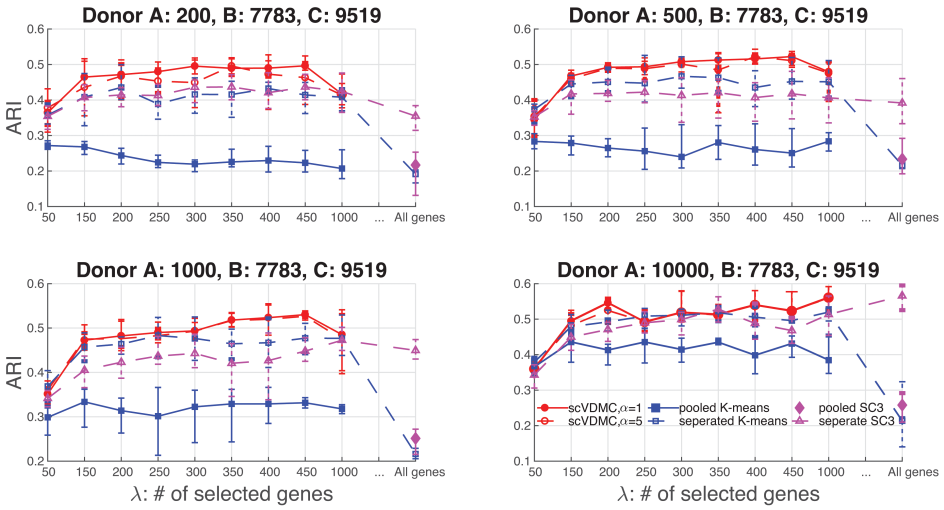


Figure 2.4: **Clustering performance on PBMC dataset.** The clustering performance of scVDMC compared with the baseline methods on the single cells from donor A measured by adjusted rand index (ARI). Pooled k-means, separated k-means, SC3 and scVDMC are tested with varying numbers of selected marker genes. Separated k-means Seurat, cellTree, SNN-Cliq and SC3 are tested using all the genes as input to the software/program. To show the strength of multitask learning, different numbers of cells, 200, 500, 1000 and 10000, are sampled from the donor A data and combined with the 7783 cells from donor B and 9519 cells from donor C for clustering. The hyper-parameters for scVDMC are $w = 0.5, \alpha = 1$ or 5.

comparable results in the original paper [139] as well as pooled and separated k-means. Both analyses were able to detect and highly rank the known marker genes of the different cell types: Clara (*Scgb1a1*), AT1 (*Pdpn, Ager*), and AT2 (*Sftpc, Sftpb*). Further, scVDMC detected several additional genes that pooled k-means, separated k-means and the original paper did not. These included two components of the Notch signaling pathway (*Notch1* and *Nrarp*) previously shown to be critical for the development of lung alveolar spaces, with AT2 cells being major sites of Notch activation [143].

2.2.9 Experiment on peripheral blood mononuclear cells data

The peripheral blood mononuclear cells (PBMC) data generated by [140] was downloaded from the 10xGenomics website. In the original data, there are 10 bead-enriched subpopulations of PBMC from a fresh donor (Donor A) with 93802 cells in total. In addition, there are also PBMC from two other frozen donors (Donor B and C) with 7783 and 9519 cells, respectively. A massive droplet-based method was applied to count the mRNAs in the tens of thousands of cells in parallel. To better evaluate the multitask learning setting, each of the 10 subpopulations of Donor A in proportion to the sizes of the populations were sampled to obtain four subsets of cells from Donor A with 200, 500, 1000 and 10000 cells by sampling. The sampling procedure was repeated five times to generate the mean and variance of ARIs. All the cells in Donor B and C were kept. Genes expressed in less than 3 cells were removed which resulted in 17647 remaining genes.

To determine the number of clusters in the PBMC data, The"elbow" plot was examined for all three cell populations shown (Supplementary Figure S6). The plots show consistent patterns in the three cell populations that the "elbow" is observed to start around $k = 10$ verifying that there are indeed around 10 cell types in the data. To tune α for scVDMC, the positions of the cluster centers across the domains were examined and are visualized by PCA (Supplementary Figure S4(E) and (F)). $\alpha = 5$ is chosen since the relative positioning of cluster centers are also relatively similar in the three domains. The baseline methods k-means and SC3 are tested on the pooled data (mixture of Donor A,B and C) and separated data (Donor A only). For SC3, the hybrid approach (consensus clustering + SVM) with its default parameters is applied on the pooled data due to the scalability issue [136]. Clustering performance is measured using Adjusted Rand index (ARI) [136] by comparing the predicted labels with the true labels from sampling the ten subpopulations of PBMC in Donor A.

Figure 2.4 shows the clustering results. Compared with pooled k-means and SC3,

scVDMC shows a consistently higher ARI with different choices of λs. scVDMC also shows a significant improvement compared with separated k-means and SC3 when there are 200, 500 and 1000 cells from Donor A. The improvement by scVDMC becomes only marginal when there are 10000 sampled from Donor A. The observation is common since larger dataset often benefit less from multitask learning, i.e. as the sample size in donor A increases, less additional information carried in the data of donor B and C can inform a better clustering of donor A data. On this dataset, the clustering performance of scVDMC does not appear to rely on the parameter α. This is likely because the agreement among the 10 clusters in the three domains is already high when $\alpha = 0$ as shown in Supplementary Figure S4(E). Therefore enforcing stronger agreement by increasing α will not lead to big improvement as shown in Supplementary Figure S4(F). Overall, scVDMC performed well on the large-scale data showing the advantage of applying multitask learning. SC3 did not over-perform separated k-means indicating the consensus clustering is less effective on this dataset.

2.2.10 RDEB scRNA-seq data

To determine the number of clusters in the RDEB data, the"elbow" plot in all the six cell populations was exampled and is shown in Supplementary Figure S7. The plots show consistent patterns in all six cell populations that the "elbow" starts from $k = 4$, which was chosen as the number clusters for clustering in all the experiments on the RDEB data. The convergence of scVDMC on RDEB data is shown in Figure S2(D).

Applying scVDMC to the RDEB single-cell dataset revealed quite different cell population structures for the six patient-sibling pairs. As shown in Figure 2.5, the agreement among the cluster centers across the six populations varies under different choices of α. When $\alpha = 0$, no agreement among the cluster centers are required. The arrangement of the four cluster centers are very different in the six populations (Figure 2.5(A)). With larger values of α, the arrangement

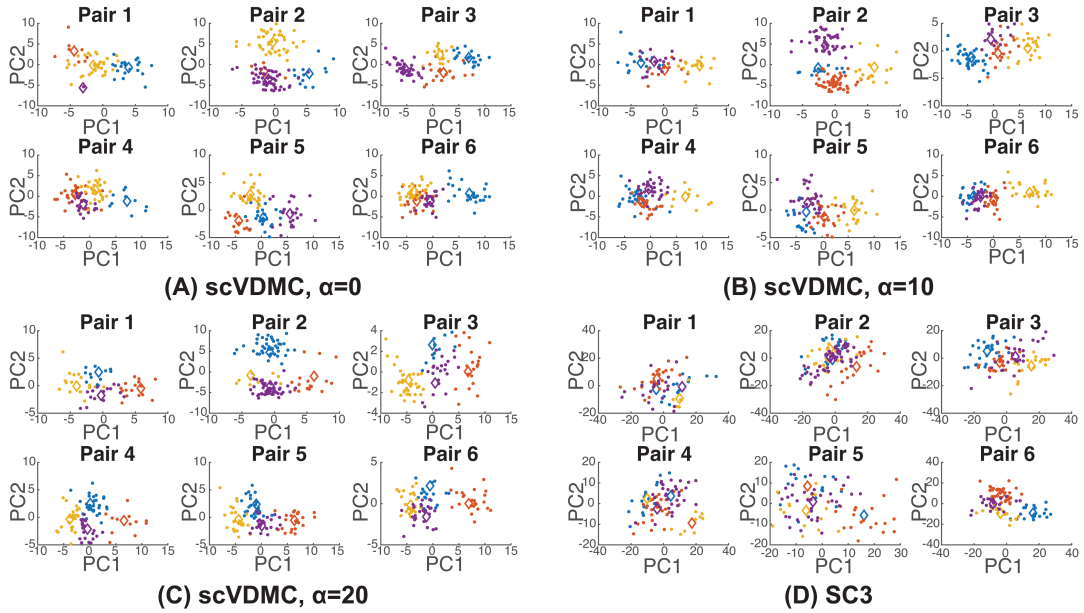


Figure 2.5: **Distinct single-cell populations from six RDEB patients and their matched siblings.** In (A), (B) and (C) PCA is applied to the combined single cell profiles of the learned marker genes by scVDMC from the six cell populations. parameters $\alpha = 0$, 10 and 20 are tested. (D) PCA is applied to the combined single cell profiles of all the genes from the six cell populations and the clusters are found by SC3 are shown. Each plot shows the projection by the first two principle components. The cluster centers are indicated by the diamonds.

of the cluster centers becomes more similar. When $\alpha = 20$, the structure of the four cluster centers is almost identical for the six populations (Figure 2.5(C)). The visualization in Figure 2.5 clearly illustrates the effect of imposing variance constraint on the cluster centers across the populations to account for the population specificity and commonality. For comparison, SC3 was also applied to the pooled cell populations and the individual cell populations. SC3 failed to detect any cluster structures in the pooled cell populations by simply clustering the cells based on the sample origin as shown in Supplementary Figure S5. SC3 also only detected inconsistent clusters across the six populations as shown in Figure 2.5(D) as expected since SC3 unlike scVDMC only clusters the cell populations

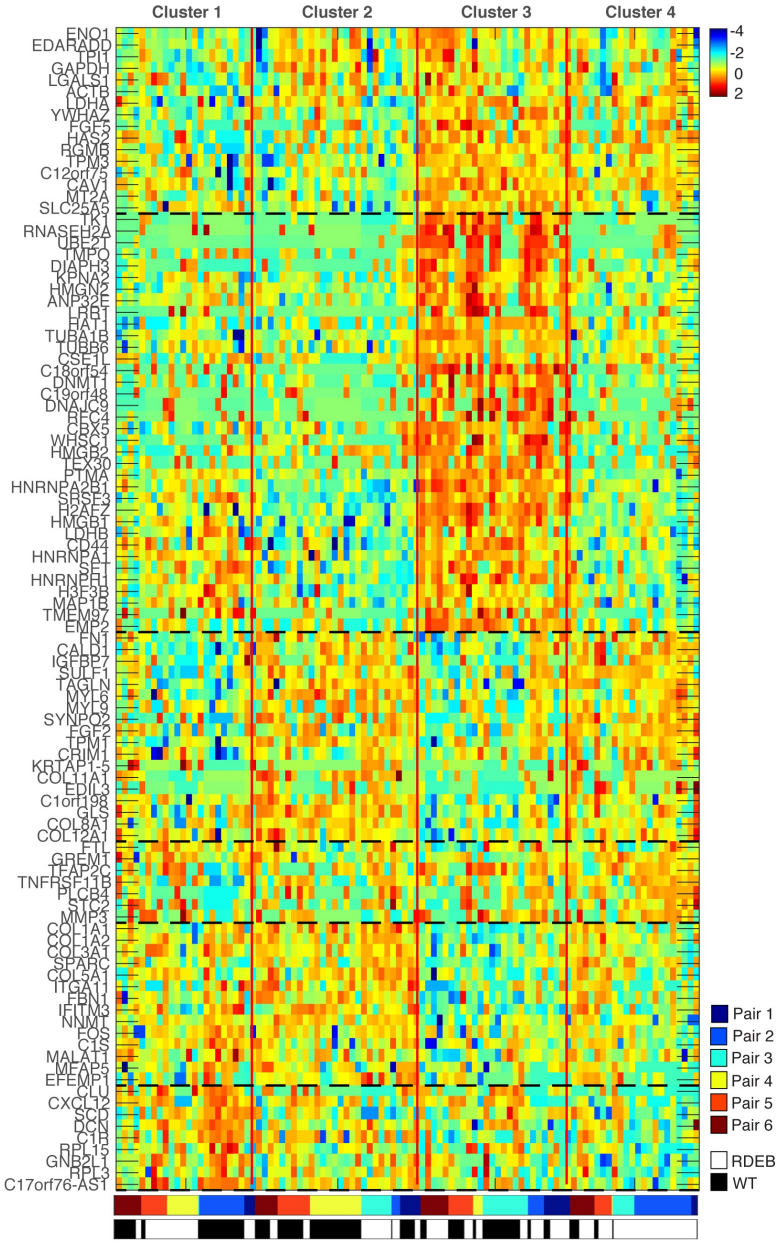


Figure 2.6: **Single-cell clustering by 100 markers genes on the RDEB data with scVDMC.** The solid vertical red lines separate the cell clusters and the black dashed horizontal lines indicate marker gene clusters derived by hierarchical clustering. The sample origin of the single cells are also annotated at the bottom by the color bars.

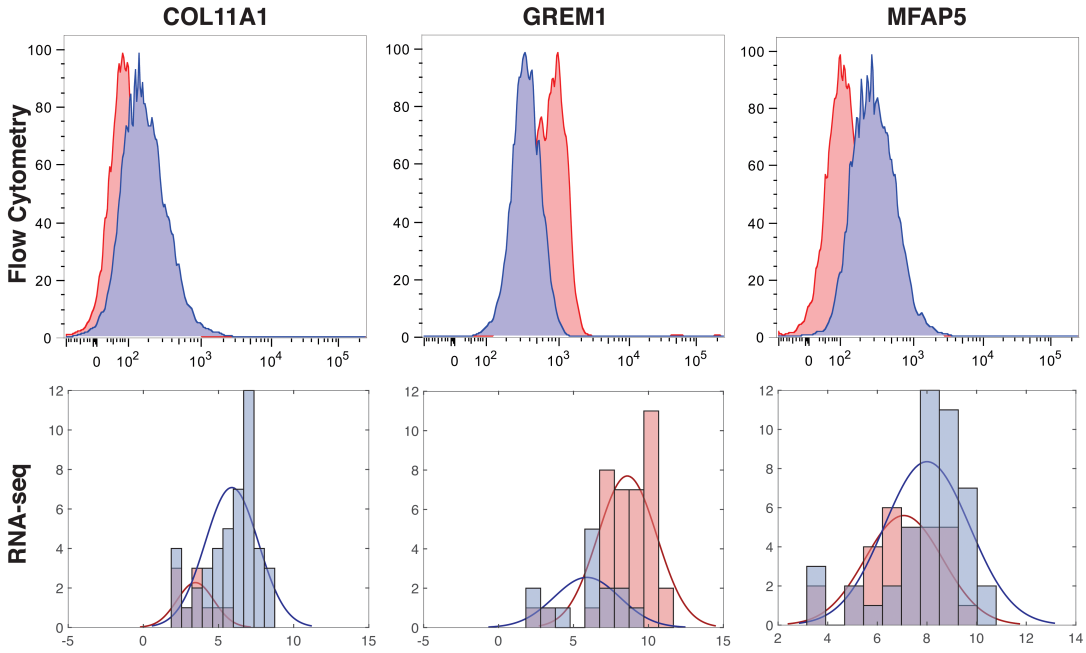


Figure 2.7: **Validation of the novel markers by flow cytometry.** The distribution of expressions for novel genes was similar between flow cytometry experiments (top) and the single-cell RNA-seq data (bottom) for the genes *COL11A1, GREM1,* and *MFAP5*. RDEB patient single-cells are shown in red; matched sibling single-cells are shown in blue. Flow cytometry data are measured as percent of max; RNA-seq data measured in FPKMs. RDEB-WT Pair 4 shown for *COL11A1* and *MFAP5*; RDEB-WT Pair 1 shown for *GREM1*.

independently.

scVDMC identified several marker genes previously known to be involved in RDEB (Figure 2.6). These included *CXCL12/SDF1*, the ligand for *CXCR4*, which directs cells of the bone marrow to damaged tissue including skin [144] and *HMGB1*, which has shown to be positively correlated with RDEB severity [145] and also mediates recruitment of bone marrow-derived cells to injured tissue [146]. Note that confounding cell cycle genes were empirically removed from the top 100 predicted markers and repeated scVDMC until there were no selected cell

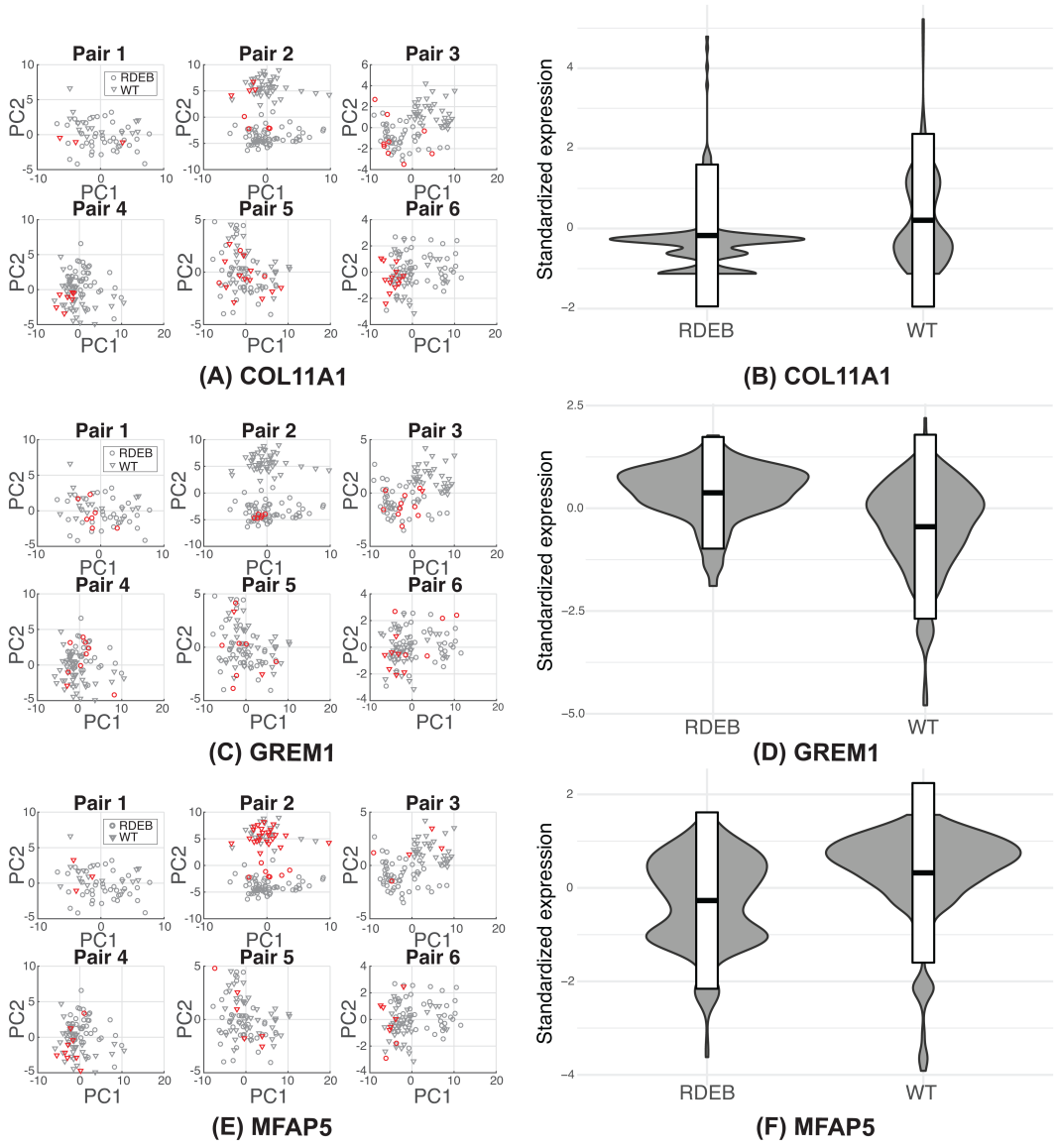


Figure 2.8: **The expressions of the markers genes in the RDEB cells and WT cells.** The scatter plots in (A), (C) and (E) show the single cell profiles of the top-100 genes projected to the first two principal components obtained by PCA with the circles representing RDEB cells and triangles representing WT cells. The cells with highly expressed markers are marked in red. The violin plots in (B), (D) and (F) show the distribution of the marker gene expressions in the RDEB cells and WT cells combined from the six pairs.

cycle genes.

Several genes were identified as markers that had not previously been associated with RDEB. These included *COL11A1*, a minor fibrillar collagen shown to mark activated cancer-associated fibroblasts (CAFs) that is not typically expressed in fibroblasts associated with inflammation and fibrosis [147]. scVDMC also revealed *GREM1*, a BMP antagonist associated with renal and pancreatic fibrosis [148, 149] and *MFAP5*, which promotes attachment of cells to micro-fibrils of the extracellular matrix and interacts with TGBβ growth factors [150]. Flow cytometry was performed on the same RDEB patient and matched sibling fibroblasts to validate the expression levels of these genes at the single-cell level and found the results similar to our RNA expression data shown in Figure 2.7. To further investigate the expressions of the these markers among the cells in the six populations, the distribution of the cells with highly expressed markers was plotted for the six pairs in Figure 2.8. In the plots, the expression patterns of *GREM1* and *MFAP5* are very consistent among the cells in all the six pairs with more enrichment in RDEB cells (*GREM1*) or WT cells (*MFAP5*). The expression pattern of *COL11A1* is consistent in five of the pairs with enrichment in WT cells except RDEB-WT pair 3. Since the markers are selected to capture cell types rather than RDEB vs WT, there might be some discrepancy in the expression patters in each individual cell populations depending on the proportion of the cell types. As top hits, these genes potentially mark sub-populations of stromal cells that contribute to the transformation of the overlying epithelium and the development of squamous cell carcinoma in RDEB patients.

2.3 Discussion

This work demonstrates that multitask learning is useful in analysis across multiple single-cell populations. It is also possible to apply other multitask learning or transfer learning methods [151] for the clustering tasks. scVDMC is a multitask clustering method specifically designed for scRNA-seq data for selection of

a smaller set of cell-type markers and allows large variability in gene expression across the cell populations. Other methods are often built using different assumptions of the data that might not be applicable to the characteristics of scRNA-seq populations [152, 153, 154].

The amount of variation across multiple scRNA-Seq datasets depends on the nature of the datasets for the integrative analysis. For example, while little variance is expected among technical replicates and slightly more variance among biological replicates such that the variances do not play a major role in the pooled analysis, much larger variance might exist among samples of different tissue types or samples from different patients as those in the RDEB data. The key hypothesis of scVDMC is the existence of a common set of a small number of marker genes in every dataset that can partition each dataset into the same clusters. While the hypothesis is independent of the amount of variation across the datasets, scVDMC formulation accounts for the variation by tuning the parameter α to weight the variances. In theory, scVDMC is applicable to the general integration of scRNA-Seq datasets if the variances calculated among the cluster centers across the datasets well represent the underlying variations. However, in real applications, it is difficult to assess if the variations are captured by the computation of the variances. Thus, more careful practice of parameter tuning and validation of the results are necessary after the application of scVDMC.

There are limitations in the scVDMC method. In multitask clustering, assuming a global k as the number of clusters in each cell population dataset does not always hold true as for some rare cell types, the corresponding cells may only be present in some populations. scVDMC might incorrectly split a cluster of one cell type because no empty cluster is allowed. One possible improvement is to model each domain with an individual $k^{(d)}$ with a more adaptive strategy for choosing $k^{(d)}$. In this case, the overall balance between within-cluster distance and the variance will need to be more carefully weighted. In addition, cell-cycle-associated genes could be a large source of confounders. Unless the stages of cell cycle are the biological signal under study, cell cycle-related variation could obscure biological

signals of interest. It is possible to model the confounders directly in the scVDMC method with more complex modeling. Alternatively, the scRNA-seq data could be pre-processed to remove the cell cycle signals. For example, a Gaussian processes-based latent-variable model [155] was used to account for confounding variations due to the cell cycle in scRNA-seq data sets and then linear regression was applied to remove them. In this approach, a clearly defined cell cycle gene set is necessary to avoid removing true signals unexpectedly. Combined with the pre-processing, scVDMC might achieve further improvement in clustering multiple cell populations.

For a better interpretation of scRNA-seq data, CellTree [134] based on Latent Dirichlet allocation also provides soft cluster assignment as opposed to the hard one-cluster assignment and more recently, a new method [156] was introduced for visualizing the cluster membership of single cells by the soft cluster assignment known as "grades of membership". It is also possible to extend scVDMC method to perform soft cluster assignment by relaxing V to contain positive real numbers rather than binary 0/1 in equation 2.2. The relaxation will require solving many least-squares problems and increase the computational time complexity. A possible future direction is investigating better solutions of scVDMC for soft cluster assignment and handling cell-cycle-associated gene signatures.

Chapter 3

Modeling and Rescue of Defective Blood-Brain Barrier Function of Induced Brain Microvascular Endothelial cells from Childhood Cerebral Adrenoleukodystrophy Patients

3.1 Hypothesis

X-linked adrenoleukodystrophy (X-ALD) is caused by mutations in the ABCD1 gene. 40% of X-ALD patients will convert to the deadly childhood cerebral form (ccALD) characterized by increased permeability of the brain endothelium that constitutes the blood-brain barrier (BBB). Mutation information and molecular markers investigated to date are not predictive of conversion. Prior reports have

focused on toxic metabolic byproducts and reactive oxygen species (ROS) as instigators of cerebral inflammation and subsequent immune cell invasion leading to BBB breakdown. This study focuses on the BBB itself and evaluates differences in brain endothelium integrity using cells from ccALD patients and wild-type (WT) controls. I hypothesized that the integrity of the blood-brain barrier of patients with cerebral childhood adrenoleukodystrophy is decreased.

In this study, we used a previously established directed differentiation protocol to derive iBMECs from WT- and ccALD-iPSCs. This enabled us to model the BBB of ccALD patients and to examine potential differences in barrier function specific to ccALD. P188 and a PEO-PPO$\underline{t}$ diblock copolymer, $E_{182}P_{16}\underline{t}$, were investigated for their potential to improve BMEC integrity. Testing of these two copolymers with this BBB model is a new avenue of investigation for X-ALD. Improvements in barrier function produced by amphiphilic block copolymers have implications for translation into a treatment for preventing the onset of ccALD by improving the BBB integrity of X-ALD patients. Translating the results from this study has the potential to reduce the number of individuals with X-ALD who develop deadly and rapidly progressive ccALD.

3.2 Materials and methods

3.2.1 Derivation and culture of hiPSCs

Normal and ccALD iPSC lines (Supplementary Table S4) were used [157, 117]. Cell lines were reprogrammed using retroviral gene delivery using the reprogramming factors OCT4, SOX2, KLF4, and c-MYC (Addgene) (WT1, WT2, ccALD1, ccALD2, ccALD3) or obtained from American Type Culture Collection (ATCC) (WT3 = ACS-1024). Cells were derived from somatic cells on irradiated MEF cultures and transferred to Matrigel (Corning) and E8 Medium (Thermo Fisher

Scientific) or TeSR-E8 (STEMCELL Technologies) for additional feeder-free expansion and maintenance. All cell lines tested negative for Mycoplasma contamination via a MycoAlert Mycoplasma Detection Kit (Lonza). With the exception of the cell line obtained from ATCC, all cell lines were authenticated using genetic fingerprinting and were also found to be karyotypically normal.

3.2.2 hiPSC differentiation to iBMECs

hiPSCs were differentiated according to Stebbins et al. (2016) [82]. On Day 8, cells were subcultured at a ratio of 1 well of a 6-well plate to 3 wells of a 12-well plate, 6 wells of a 24-well plate, 3 Transwell filters (12 mm), or 11.4 wells of a μ-slide. When the cells reached confluence 48 h after subculture (Day 10), cells were utilized for permeability, efflux transporter, immunocytochemistry, RT-PCR, RNA-seq, and Oil-Red-O staining experiments using endothelial cell (EC) medium: human endothelial serum free medium (Thermo Fisher Scientific) with 1% platelet-poor plasma derived serum (Biomedical Technologies).

3.2.3 Immunocytochemistry

iBMECs were subcultured onto μ-Slides (Ibidi). 48 h post-subculture, cells were fixed in ice-cold 100% methanol (MilliporeSigma). Fixed iBMECs were blocked for 1 h at room temperature in a blocking buffer of PBS containing 10% normal goat serum (Thermo Fisher Scientific). Cells were incubated with primary antibodies diluted in blocking buffer overnight at 4°C. After 3 washes with PBS for a minimum of 5 minutes per wash, cells were incubated with secondary antibody for 1 h in the dark at room temperature (Supplementary Table S5, Supplementary Table S6). Cells were subsequently washed with PBS three times for a minimum of 5 minutes per wash and incubated with 4', 6-diamidino-2-phenylindoldihydrochloride (DAPI; Thermo Fisher Scientific) for 5 minutes to label nuclei. Cells were washed once with PBS for 5 minutes before imaging with an EVOS FL Auto Cell Imaging microscope.

3.2.4 RT-PCR

Cells were differentiated as described above and detached with trypsin. Total RNA was extracted using an RNeasy Mini Kit (Qiagen) following the manufacturer's protocol and quantified using a NanoDrop®ND-1000. cDNA was generated from 1 g of RNA using Omniscript reverse-transcriptase (Qiagen) and oligo-dT primers (Thermo Fisher Scientific). RT-PCR was performed using the GoTaq Green Master Mix (Promega) and PrimePCR primer sets (Bio-Rad) (Supplementary Table S7). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as the housekeeping gene. Gel electrophoresis of RT-PCR products with a 2% agarose gel was used to analyze transcript amplification.

3.2.5 Trans-endothelial electrical resistance

iBMECs were seeded onto Transwell filters. TEER was measured daily starting 24 h after subculture utilizing the EVOM2 voltohmmeter with STX3 chopstick electrodes (World Precision Instruments). TEER was measured on an empty Transwell filter coated with collagen and fibronectin, and this value was subtracted from the TEER of the cell monolayer each time. TEER values were normalized by the surface area of the Transwell filter.

3.2.6 Sodium fluorescein permeability

iBMECs were seeded onto Transwell filters. An empty Transwell filter coated with collagen and fibronectin was utilized to measure the permeability of the membrane. After a complete medium change, the cells were incubated at 37°C for 1.5 h. TEER was measured before and after the medium change to confirm monolayer equilibration. Medium from the apical chamber was aspirated and replaced with EC medium containing 10 μM sodium fluorescein (MilliporeSigma). Every 30 min for 2 h, 150 μL aliquots were extracted from the basolateral chamber and replaced with 150 μL of fresh medium. At 2 h, a 150 μL sample was extracted from the apical chamber and then fluorescence was measured on a BioTek Synergy

H1 multi-mode microplate reader at excitation of 485 nm and emission of 530 nm. Calculation of sodium fluorescein permeability was done following Stebbins et al. (2016) [82].

3.2.7 Rhodamine 123 accumulation

Accumulation of rhodamine 123, a P-glycoprotein (P-gp) substrate, was measured in the absence and presence of a P-gp inhibitor cyclosporin A to quantify P-gp efflux potential. iBMECs were seeded onto 24-well plates. Cells were pre-incubated with or without 10 μM cyclosporin A (MilliporeSigma) in HBSS (Thermo Fisher Scientific) for 1 h at 37°C. Next, all cells were incubated with 10 μM rhodamine 123 (MilliporeSigma) in HBSS for 2 h at 37°C. Following the incubation steps, cells were lysed using RIPA buffer (MilliporeSigma) and fluorescence was measured on a BioTek Synergy H1 multi-mode microplate reader at excitation of 485 nm and emission of 530 nm. Unlysed cells from a parallel setup were dissociated with Accutase (Thermo Fisher Scientific) and counted using the Countess II to normalize the fluorescence on a per cell basis.

3.2.8 Analysis of tight junction continuity

For quantitative analysis of iBMEC integrity, the percentage of cells expressing frayed tight junctions was counted using iBMECs immunolabeled for occludin. Cells were defined as having frayed tight junctions if any cell-cell contact point appeared discontinuous. A blinded analysis in which three different people each counted fifteen separate frames and 12,530 total junctions was used to obtain a percentage of frayed tight junctions for both the ccALD- and WT-iBMECs.

3.2.9 Electron microscopy

iBMECs were seeded onto 6-well plates. Two days after subculture, the cells were fixed with 1 mL of 2.5% glutaraldehyde (Electron Microscopy Sciences) in 0.1 M

sodium cacodylate for 1 h at room temperature. Using cell lifters to detach the cells while preserving cell-cell junctions, cells were collected in microcentrifuge tubes and stored in fresh fixative solution for pelleting via centrifugation. Following 3 washes with 0.1 M sodium cacodylate buffer, cells were post-fixed with 1% osmium tetroxide (Electron Microscopy Sciences). Cells were dehydrated in acetone and subsequently embedded with Embed 812 resin (Electron Microscopy Sciences). A Leica UC6 Ultramicrotome (Leica Microsystems) was used to section the embedded samples. A JEM 1400 Plus transmission electron microscope (JEOL LTD) and AMT Capture Engine Version 7.00 (Advanced Microscopy Techniques Corp.) were used to analyze and image the samples.

3.2.10 RNA-sequencing

Cells were differentiated as described above and detached with trypsin. Total RNA was isolated from WT- and ccALD-iBMECs using an RNeasy Mini Kit (Qiagen) following the manufacturer's protocol. RNA with a RNA Integrity Number (RIN) score > 8 was used for library generation with the TruSeq Stranded mRNA Sample Preparation kit (Illumina). Paired-end 150 bp length reads were generated using an Illumina MiniSeq. Low quality bases were trimmed using Trimmomatic (enabled with the optional "–qualitycontrol" option and a 3bp sliding-window trimming from the 3' end requiring minimum Q16). The remaining reads were mapped to hg19 using Tophat2. The featureCounts program in the R SubRead package was used to generate a transcript abundance file for input into the R package edgeR to identify differentially expressed genes. Ingenuity Pathway Analysis [158] was used for network analysis and Gene Ontology [159] for pathway analysis.

3.2.11 Oil-Red-O staining and image analysis

iBMECs were seeded onto 12-well plates. Cells were fixed with 10% formalin for 20 minutes, subsequently dehydrated with 60% isopropanol, and incubated with Oil-Red-O (MilliporeSigma) for 10 minutes before being washed 4 times with

deionized water. Images were captured using a Nikon Eclipse TS100 inverted light microscope connected to a Unitron Microscopes Lumenera®Cameras AU-310-CMOS Infinity 1 camera. To measure the abundance of lipid droplets from the Oil-Red-O stained images, a custom MATLAB script was used to quantify the number and intensity of red pixels. Red pixels were defined on the HSV (hue, saturation, value) scale as having hue between 0.833 and 0.073, saturation between 0.300 to 1, and value between 0 and 1. The MATLAB Color Thresholder tool was used to mask as black any pixels not defined as red. Average intensity was measured by summing the value in the red channel of the RGB scale for each pixel, while the average number of pixels was calculated as the total number of red pixels.

3.2.12 Diblock copolymer synthesis

$E_{182}P_{16}\underline{t}$ diblock copolymer was synthesized via ring-opening anionic polymerization following established techniques necessary for air and water free environments described elsewhere [100, 160, 161] with alumina column-dried tetrahydrofuran (THF) as the solvent. The PPO block was first synthesized at room temperature by initiation with potassium tert-butoxide (MilliporeSigma) in the presence of 18-crown-6 ether (MilliporeSigma) [162, 163, 164]. The reaction was carried out for 48 h, after which the reaction was terminated with excess acidic methanol (1:10 w/w% hydrochloric acid/methanol) to give tert-butoxy-terminated PPO chains. The PPO homopolymer was purified by iterative filtration, solvent removal, and dissolution in fresh THF. Subsequently, the hydroxyl terminated PPO was reinitiated with potassium naphthalenide, reacted with ethylene oxide for 20 h, and then terminated with excess acidic methanol. The resulting diblock copolymer was purified by iterative filtration, solvent removal, and dissolution in fresh THF. The final product was retrieved upon an additional purification step via dialysis.

3.2.13 Polymer characterization

P188 (MilliporeSigma) and $E_{182}P_{16}\underline{t}$ were characterized via proton nuclear magnetic resonance spectroscopy (Bruker AX-400; deuterated chloroform as solvent) to determine compositions and/or number average molecular weight by end-group analysis. A size exclusion chromatograph (Waters) with a refractive index detector was used to obtain dispersity of the polymers. THF was utilized as the solvent, and the chromatograph was calibrated with polystyrene standards. The weight percent of PEO and dispersity of P188 were found to be 80 wt% and 1.06, respectively (number average molecular weight of P188 was provided by the manufacturer to be 8400 g/mol). The weight percent of PEO, number average molecular weight, and dispersity of $E_{182}P_{16}\underline{t}$ were determined to be 90 wt%, 8900 g/mol, and 1.07, respectively (Supplementary Figure S8).

3.2.14 Polymer treatment

Working solutions of polymers were prepared by dissolving P188 and $E_{182}P_{16}\underline{t}$ in 1X DPBS without magnesium or calcium (Thermo Fisher Scientific) to a concentration of 12 mM and sterilized via filtration. For polymer treatment during development, the polymers were added to the culture on Day 3 of the differentiation protocol. The polymer working solutions were diluted to a concentration of 0.5 mM or 1 mM in the culture medium and the resulting solution was added to the culture during standard medium change. The control received medium with DPBS such that all conditions had the same volume of DPBS in the medium. After 24 h, the medium was changed in accordance with the differentiation protocol, effectively removing any excess polymers. On Day 8, the iBMECs were subcultured onto Transwells for TEER measurements or 12-well plates for Oil-Red-O staining. For polymer treatment post-differentiation, iBMECs were subcultured onto Transwells on Day 8; on Day 9, a small aliquot of the P188 or $E_{182}P_{16}\underline{t}$ working solutions was added to the apical chamber such that the final concentration of polymers in the apical chamber was 1 mM. A corresponding volume of DPBS

was added to the control cells.

3.2.15 Statistical analysis

Data are presented as mean standard error (SEM) with n defined in figure legends. P-values were determined using an unpaired Student's t-test. Statistical analysis was performed using GraphPad Prism.

3.3 Results

3.3.1 Directed differentiation of WT- and ccALD-iPSCs into iBMECs

A previously published protocol [82] was used to direct the differentiation of iPSCs into iBMECs from three clinically confirmed cases of ccALD and three WT controls. Immunofluorescence and RT-PCR demonstrated that both patient and control iBMECs expressed the requisite endothelial markers PECAM-1 and VE-cadherin (*CDH5*), the tight junction markers claudin-5 and occludin, and the BBB markers P-glycoprotein and GLUT-1 (*SLC2A1*) (Figure 3.1(a), (b)), Supplementary Figure S9. Since expression of *ABCB1*, which encodes the BMEC-specific efflux transporter P-gp, was decreased for one of the ccALD-iBMEC lines (Supplementary Figure S10(a)), we employed a rhodamine 123 accumulation assay to check the P-gp efflux potential of the ccALD-iBMECs (Supplementary Figure S10(b)). Normalized accumulation after inhibiting P-gp with cyclosporin A (CsA) was lower for the same iBMEC line in which we noticed the decreased *ABCB1* expression (102.7 ± 2.3) compared to the other ccALD-iBMEC lines (152.7 ± 23.4 and 167.1 ± 19.0) as well as the WT-iBMECs lines (258.2 ± 16.2, 220.8 ± 26.4, and 204.7 ± 20.2).

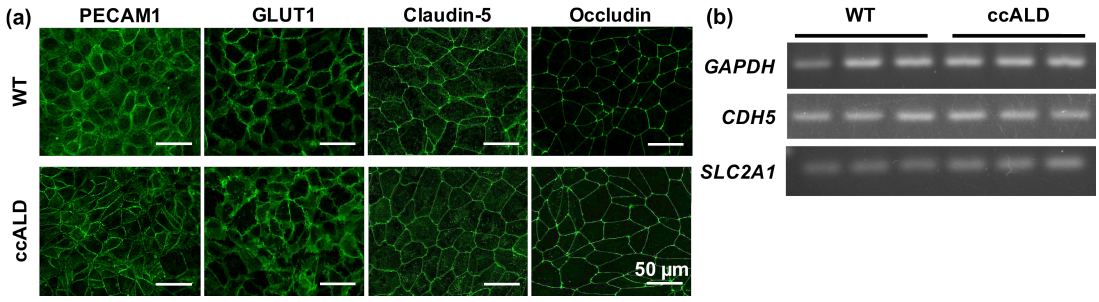


Figure 3.1: **iBMECs express the requisite endothelial, tight junction, and BBB markers. (a)** Representative immunocytochemistry (WT1 and ccALD2). iBMECs from ccALD patients and WT controls express PECAM1, GLUT1, claudin-5, and occludin. No qualitative difference was observed between the WT and ccALD-iBMECs. **(b)** RT-PCR (all WT and ccALD lines). iBMECs from ccALD patients and WT controls express *CDH5* (VE-cadherin) and *SLC2A1* (GLUT1).

3.3.2 ccALD-iBMECs have impaired barrier properties

To investigate functional differences between the ccALD- and WT-iBMECs, we used trans-endothelial electrical resistance (TEER) to measure the barrier integrity of the iBMECs on Days 1-4 following subculture onto Transwell filters. At all days measured, we found a statistically significant difference ($p < 0.0001$) in TEER between the ccALD- and WT-iBMECs (peak TEER on Day 2 of measurement: $2592 \pm 110\ \Omega \cdot cm^2$ compared to $5001 \pm 172\ \Omega \cdot cm^2$ for the ccALD-iBMECs and WT-iBMECs, respectively) (Figure 3.2(a)) (see Supplementary Figure S11 for TEER measurements for individual cell lines). Additionally, permeability of sodium fluorescein was measured to be $1.85 \pm 0.19 \times 10^{-5}$ cm/min for the ccALD-iBMECs and $1.50 \pm 0.31 \times 10^{-5}$ cm/min for the WT-iBMECs (Figure 3.2(b)). The difference in permeability is not statistically significant despite the substantial difference ($\sim 2400\ \Omega \cdot cm^2$) in TEER between the WT- and ccALD-iBMECs; however, our results are consistent with previous studies that report sodium fluorescein permeability values on the order of 10^{-5} cm/min for BMECs with TEER greater than $2000\ \Omega \cdot cm^2$ [165, 166, 82, 83] and with reports that demonstrate that

small molecule passive permeability does not correlate strongly with TEER above certain TEER thresholds [167, 168, 169, 166]. To examine potential differences in tight junction organization, we employed a frayed junction analysis. The ccALD-iBMECs had more frayed junctions ($p < 0.01$) compared to the WT-iBMECs (37 $\pm$ 3% versus 25 $\pm$ 3%) (Figure 3.2(c), (d)). Overall, iBMECs from ccALD patients appear to form a less intact cellular barrier that permits increased passive transport of ions as well as small molecules. This defect in barrier integrity may result from mislocalization of tight junction proteins between cells.

3.3.3 Lipid droplets accumulate in ccALD-iBMECs

To assess any structural differences between the ccALD-iBMECs and WT controls, we performed transmission electron microscopy (TEM) of ccALD- and WT-iBMECs using cross-sections of fixed and pelleted cells. Numerous and large lipid droplets were present in the ccALD-iBMECs, with fewer and smaller lipid droplets in the WT-iBMECs (Figure 3.3(a)) (Supplementary Figure S12). To quantify the abundance of lipid droplets in the iBMECs, Oil-Red-O staining was used. This histological stain is specific to neutral lipids and does not stain the polarized phospholipids of the cell membrane (Mehlem2013). Lipid droplets are stained bright red, and image analysis can be used to quantify either the amount (total number of red pixels) or intensity (redness of the red pixels) of red in micrographs. Quantification of the lipid deposition in ccALD- and WT-iBMECs revealed a significant increase ($p < 0.005$) in lipid abundance in the ccALD-iBMECs compared to the WT-iBMECs (Figure 3.3(b), (c)). The average intensity of red pixels in images of Oil-Red-O stained WT-iBMECs was calculated to be 1.8 $\pm$ 0.5 x 10^6 compared to 5.4 $\pm$ 1.0 x 10^6 for the ccALD-iBMECs. The average number of red pixels was calculated to be 1.1 $\pm$ 0.2 x 10^4 for the WT-iBMECs and 2.8 $\pm$ 0.4 x 10^4 for the ccALD-iBMECs (Figure 3c). Very few lipid droplets were seen in the iPSCs and no statistical difference was observed between the ccALD-iPSCs and WT-iPSCs in the number of red pixels (22 $\pm$ 8 and 35 $\pm$ 11, respectively) or the

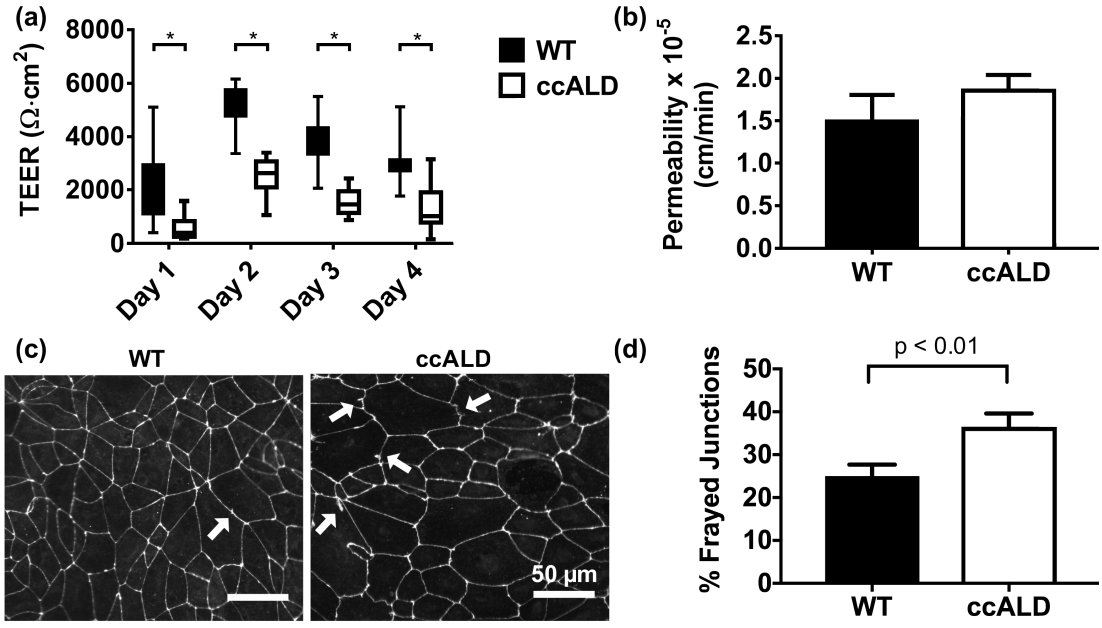


Figure 3.2: **ccALD-iBMECs are functionally distinct from WT-iBMECs.** **(a)** Trans-endothelial electrical resistance (TEER) is significantly decreased in the ccALD-iBMECs compared to the WT-iBMECs at all experimental time points. Data compiled from three independent experiments with nine biological replicates each (all iBMEC lines used) (n = 27). * $p < 0.0001$. **(b)** Passive transport as measured by sodium fluorescein permeability is slightly increased in the ccALD-iBMECs compared to WT-iBMECs. All iBMEC lines tested with three biological replicates each (n = 9). **(c)** Examples of frayed junctions indicated by white arrows on occludin immunolabeled images of WT1- and ccALD3-iBMECs. **(d)** Quantification of percent frayed junctions in WT1- and ccALD3-iBMECs indicates that WT-iBMECs have fewer frayed junctions than ccALD-iBMECs. Results of nine biological replicates with five technical replicates each shown (n = 45).

intensity of red pixels ($4.2 \pm 2 \times 10^3$ and $6.4 \pm 2 \times 10^3$, respectively) (Supplementary Figure S13). VLCFA accumulation was not observed in the ccALD-iBMECs via TEM. The presence of an increased amount of lipid droplets in the ccALD-iBMECs compared to the WT-iBMECs that arises upon differentiation (i.e. is not present during the iPSC stage) is a difference that potentially contributes to the decreased barrier integrity of the ccALD-iBMECs.

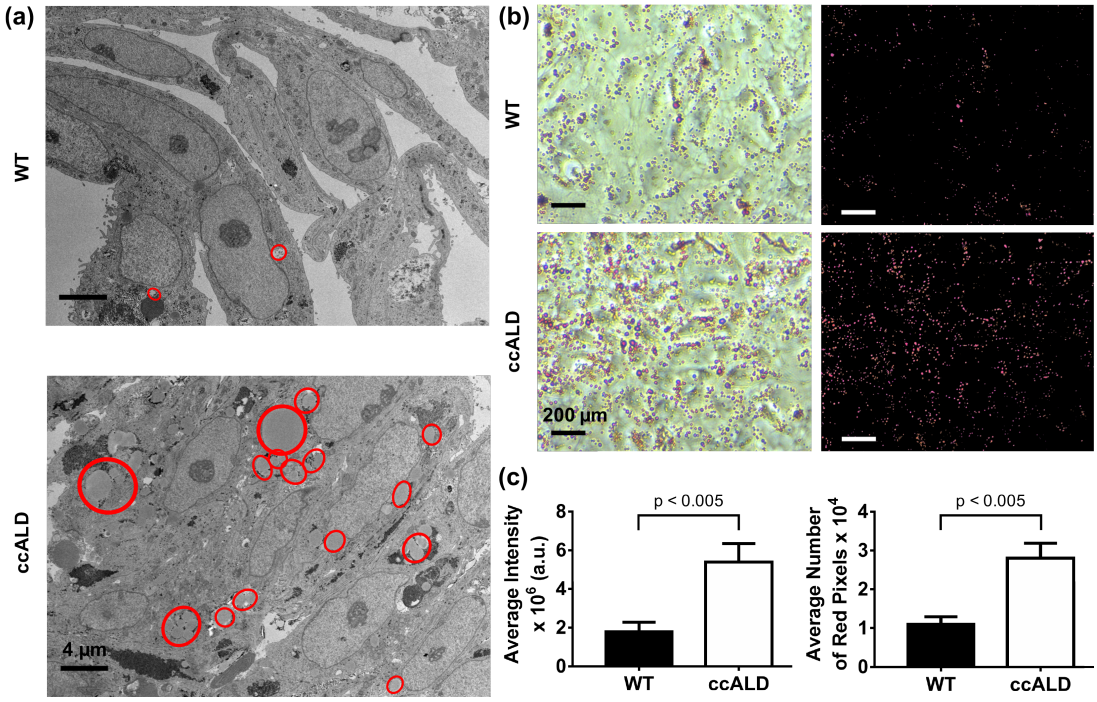


Figure 3.3: **ccALD-iBMECs accumulate more lipid droplets than WT-iBMECs.** **(a)** Comparison of transmission electron micrographs of WT1- and ccALD3-iBMECs show increased lipid droplet accumulation in ccALD-iBMECs. Lipid droplets outlined in red. **(b)** Representative images of Oil-Red-O stained WT3 and ccALD1-iBMECs. Raw images on left and masked images on right. **(c)** Quantification of intensity and number of red pixels in images of Oil-Red-O stained iBMECs indicate increased lipid droplet accumulation in ccALD-iBMECs compared to WT-iBMECs. Oil-Red-O staining images of all iBMEC lines were used for quantification using three biological replicates for each cell line (n = 9).

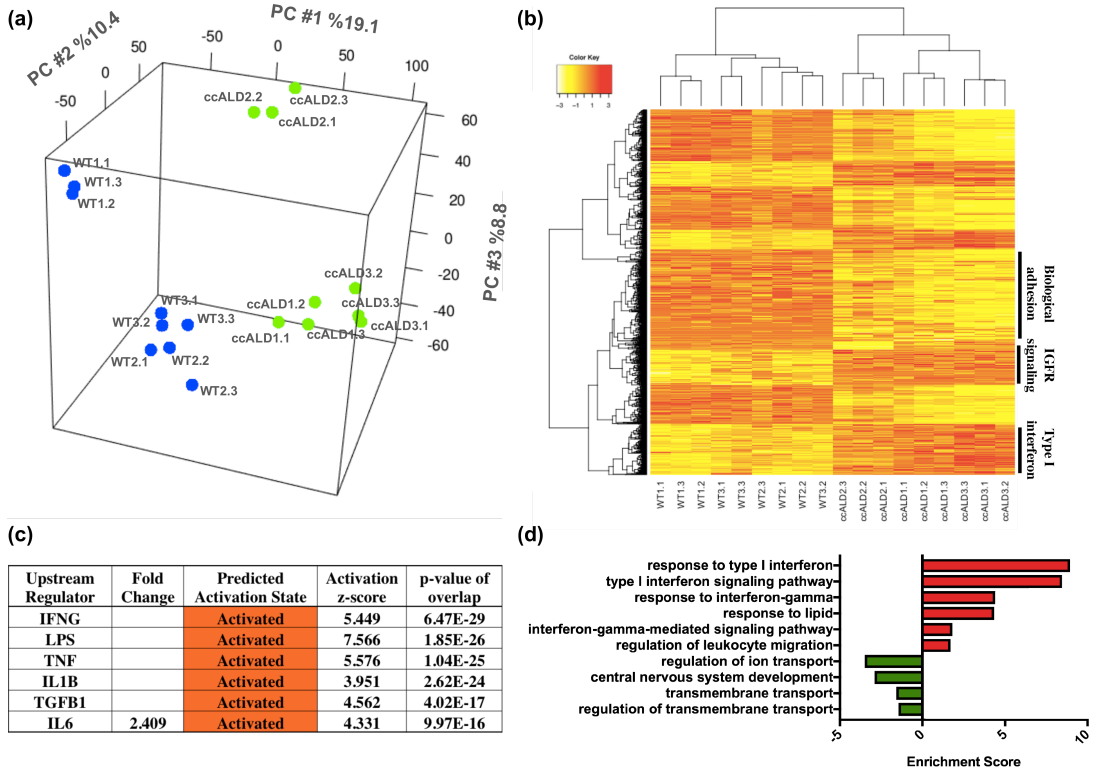
3.3.4 Transcriptome analysis indicates differences in Type I interferon activation and lipid metabolism pathways

To further characterize differences between ccALD- and WT-iBMECs and to elucidate potential mechanisms for the decreased barrier integrity seen in the ccALD-iBMECs, we performed RNA-sequencing of three replicate differentiations for each of our ccALD- and WT-iBMEC lines. Principal component analysis (PCA) separated the WT-iBMECs from the ccALD-iBMECs along the first principal component (Figure 3.4(a)). Hierarchical clustering of differentially expressed genes (DEGs) (2X fold change, false-discovery rate (FDR) < 0.05) and samples revealed a cluster of genes that were decreased in the ccALD-iBMECs involving the attachment of cells to each other including intracellular attachment between membrane regions (Gene ontology (GO): 0022610), while Type I interferon-activated signaling (GO: 0060337) and insulin-like growth factor receptor signaling (GO: 0043568) pathways were increased in the ccALD-iBMECs (Figure 3.4(b)) [170]. We used Ingenuity Pathway Analysis (IPA) to query upstream regulators of our DEGs [158]. IPA builds a graph-based network from gene expression data and uses this information to predict upstream regulators. The z-score (calculated as the number of standard deviations from the mean of a normal distribution of activity edges using this graph-based network) represents the magnitude of bias in gene regulation that predicts the activity of specific upstream regulators. This analysis revealed upstream regulators involving TGFβ1 signaling, Type I interferon response, and other immune signaling signatures highly activated in the ccALD-iBMECs (IFNG, LPS, TNF, and TGFβ1 z-scores of 5.5, 7.6, 5.6, and 4.6 respectively) (Figure 3.4(c)). GO analysis was performed on genes differentially expressed between the ccALD- and WT-iBMECs. This analysis calculates a p-value based on enrichment of genes in a particular GO annotation (-$\log_{10}$(p-value) reported as enrichment score for upregulated genes, $\log_{10}$(p-value) reported as enrichment score for downregulated genes). GO analysis on genes upregulated (increased activity) in

ccALD-iBMECs indicated an increase in Type I interferon signaling (enrichment score 8.45) and response to lipid pathways (enrichment score 4.34). GO analysis on genes downregulated (decreased activity) in ccALD-iBMECs indicated a decrease in transmembrane and ion transport (enrichment scores -1.58 and -3.5, respectively) (Figure 3.4(d)). The lipid pathway upregulation is consistent with both the primary ccALD phenotype and our TEM results, while Type I interferon signaling and other inflammatory pathways are secondary and could have many sources.

3.3.5 Block copolymers reverse impaired barrier integrity and mitigate lipid accumulation

We next investigated whether polymer treatment could rescue the impaired barrier integrity of the ccALD-iBMECs. ccALD-iBMECs were treated with 1 mM of P188 or $E_{182}P_{16}\underline{t}$ at the end of the differentiation protocol (Day 9) or during development (Day 3) (see Figure 3.5(a) for chemical structures of the polymers). Day 3 was chosen because the cells begin to express endothelial cell markers at this time point (Lippmann2012). Polymer treatment with either P188 or $E_{182}P_{16}\underline{t}$ at the end of the differentiation protocol showed minimal effect on ccALD-iBMEC TEER (Supplementary Figure S14(a)). However, we saw a significant effect ($p < 0.05$) when the ccALD-iBMECs were treated with the diblock copolymer ($E_{182}P_{16}\underline{t}$) during development. The maximum TEER of the ccALD-iBMECs treated with $E_{182}P_{16}\underline{t}$ was $3316 \pm 246\ \Omega \cdot cm^2$. This was higher than both the untreated and P188 treated ccALD-iBMECs ($2409 \pm 254\ \Omega \cdot cm^2$ and $2162 \pm 260\ \Omega \cdot cm^2$, respectively; see Figure 3.5(b)). The effect of dosage was investigated by treating the ccALD-iBMECs with 0.5 mM or 1 mM $E_{182}P_{16}\underline{t}$ on day 3 of the differentiation protocol, and we observed a larger increase in TEER compared to the control when treated with 1 mM $E_{182}P_{16}\underline{t}$ than with 0.5 mM $E_{182}P_{16}\underline{t}$ (Supplementary Figure S14(b), Additional File 1). Notably, the barrier function of the WT-iBMECs was unaffected by polymer treatment as the TEER of the untreated WT-iBMECs



(c)

Upstream Regulator	Fold Change	Predicted Activation State	Activation z-score	p-value of overlap
IFNG		Activated	5.449	6.47E-29
LPS		Activated	7.566	1.85E-26
TNF		Activated	5.576	1.04E-25
IL1B		Activated	3.951	2.62E-24
TGFB1		Activated	4.562	4.02E-17
IL6	2.409	Activated	4.331	9.97E-16

Figure 3.4: **Transcriptome analysis indicates differences in Type I interferon activation and lipid metabolism pathways.** **(a)** PCA mapping of $\log_2$ normalized read counts on global gene expression. The first three dimensions account for 38.3% of the total variance with grouping of individual WT- and ccALD-iBMEC replicates and separation of the experimental and control samples along PC1. **(b)** Heat map of DEG (n=1381) on log2 normalized read counts. Cluster annotations are from Gene Ontology analysis. **(c)** IPA upstream regulator analysis of transcriptional regulators predicted by activation z-scores. p-values calculated by Fisher's exact test using expected and observed genes overlapping with the WT versus ccALD DEGs and all genes regulated by each transcriptional regulator. **(d)** GO terms of pathways upregulated in ccALD in red with downregulated pathways in green. Data analyzed from three independent experiments with three biological replicates each (n = 9).

($3074 \pm 127\ \Omega \cdot cm^2$) was not significantly different than that of WT-iBMECs treated with P188 ($2998 \pm 50\ \Omega{\cdot}cm^2$) or $E_{182}P_{16}\underline{t}$ ($3165 \pm 95\ \Omega{\cdot}cm^2$). Additionally, lipid droplet accumulation was decreased in ccALD-iBMECs treated with 1 mM $E_{182}P_{16}\underline{t}$ during development (Figure 3.6). Quantification of Oil-Red-O staining images indicated a statistically significant ($p < 0.05$) decrease in ccALD-iBMECs treated with 1 mM $E_{182}P_{16}\underline{t}$ when compared to untreated ccALD-iBMECs. The average intensity of red pixels in images of Oil-Red-O stained ccALD-iBMECs treated with 1 mM $E_{182}P_{16}\underline{t}$ was 10.9 ± 1 x 10^6 compared to 14.8 ± 1 x 10^6 for the control. The average number of red pixels was calculated to be 4.9 ± 0.5 x 10^4 for ccALD-iBMECs treated with 1 mM $E_{182}P_{16}\underline{t}$ and 6.7 ± 0.5 x 10^4 for the control.

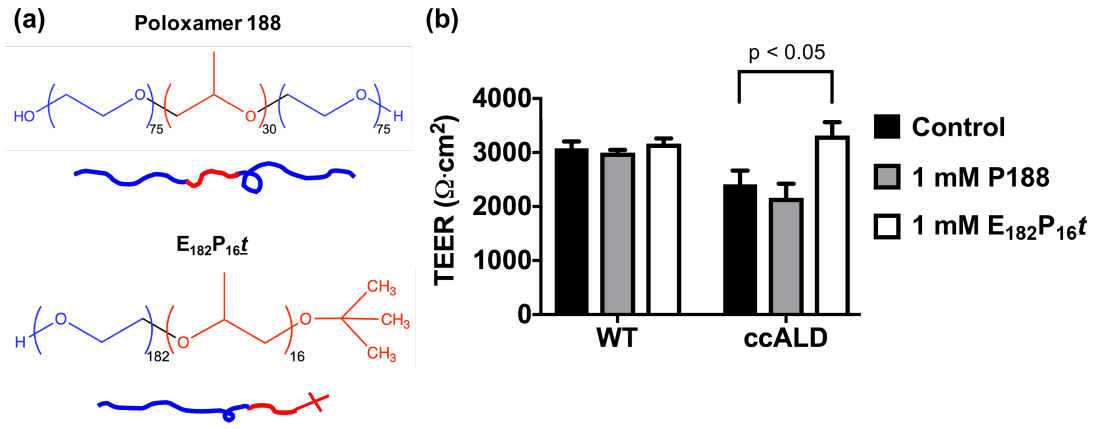


Figure 3.5: **Diblock copolymer treatment rescues defective barrier function of ccALD-iBMECs.** **(a)** Chemical structures of polymers utilized for treatment. Poloxamer 188 is a triblock copolymer of poly(ethylene oxide) (PEO) and poly(propylene oxide) (PPO); $E_{182}P_{16}\underline{t}$ is a diblock copolymer of PEO and PPO with a tert-butoxy end group on the PPO block. **(b)** Maximum TEER of WT1- and ccALD3-iBMECs treated with 1 mM of P188 or $E_{182}P_{16}\underline{t}$. Treatment with 1 mM $E_{182}P_{16}\underline{t}$ resulted in improved ccALD-iBMEC barrier function. Data shown from four independent experiments with three biological replicates each (n = 12).

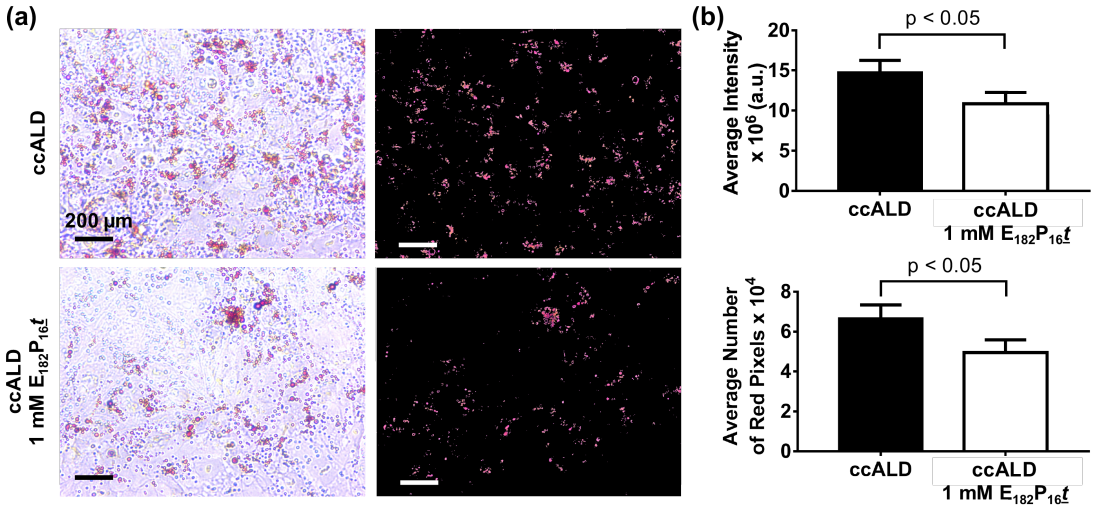


Figure 3.6: **Diblock copolymer treatment decreases lipid droplet accumulation in ccALD-iBMECs. (a)** Representative Oil-Red-O staining of untreated control ccALD3-iBMECs and 1 mM $E_{182}P_{16}\underline{t}$ treated ccALD3-iBMECs during development. Raw images shown on left and masked images on right. **(b)** Quantification of intensity and number of red pixels in Oil-Red-O stained images indicates decreased lipid droplet accumulation in ccALD-iBMECs treated with 1 mM $E_{182}P_{16}\underline{t}$ during development. Six biological replicates used for quantification (n = 6).

3.4 Discussion

Our model of the BBB demonstrating that the barrier function is defective and lipid droplets accumulate in iBMECs from patients with ccALD opens the door to new therapeutic avenues aimed at maintaining the integrity of the BBB and preventing the onset of ccALD. In the present work, our findings indicate a significant improvement in barrier function and a decrease in lipid droplet accumulation when ccALD-iBMECs are treated during differentiation with a diblock copolymer with a hydrophobic tert-butoxy end group ($E_{182}P_{16}\underline{t}$). No effect was seen when the ccALD-iBMECs were treated at the same time with P188 or when the ccALD-iBMECs were treated post differentiation with either polymer. Treatment

of WT-iBMECs with either polymer during development did not improve barrier function. Response of the ccALD-iBMECs but not the WT-iBMECs to polymer treatment further highlights that there are fundamental differences between the ccALD- and WT-iBMECs (Figure 3.7).

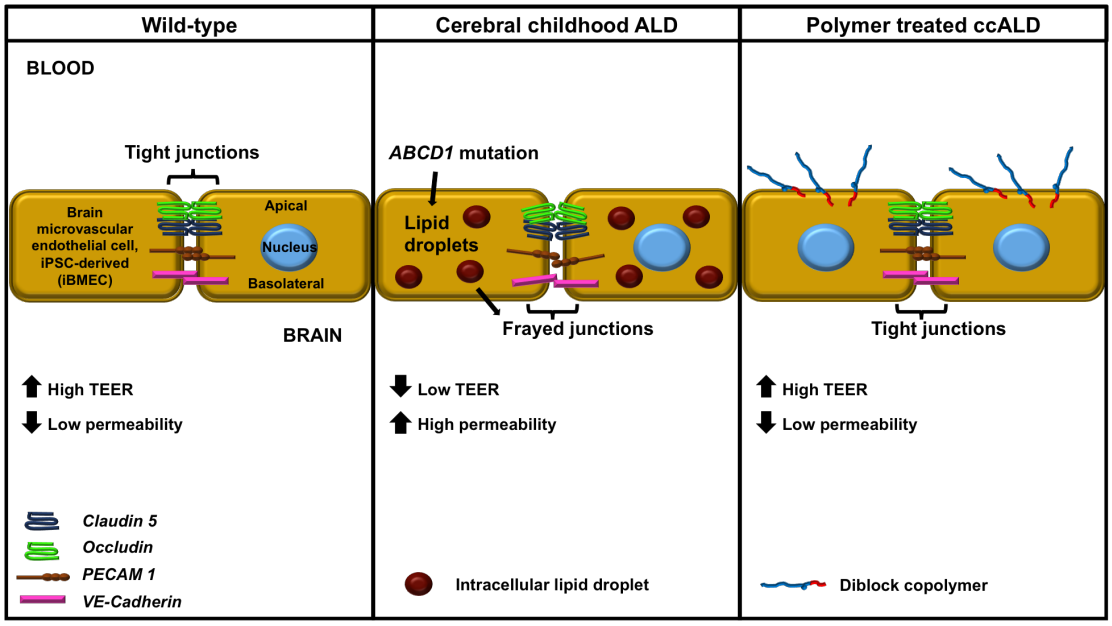


Figure 3.7: **Summary of experimental findings.** iBMECs from patients with ccALD form a less intact barrier than WT individuals. They also accumulate lipid droplets to a greater extent. Treatment with diblock copolymer $E_{182}P_{16}\underline{t}$ mitigates both of these defects.

Overall, our study demonstrates that one of the intrinsic defects in ccALD is with the integrity of the BMECs that constitute the BBB. These findings are in line with the study by Musolino et al. (2015) in which they knocked down *ABCD1* in BMECs and saw mislocalization of the tight junction protein claudin-5 [75]. The presence of frayed or discontinuous junctions and its relation to barrier function has been noted in other systems as well. In an *in vitro* epithelium model, an induced opening of the barrier for drug delivery purposes was marked by both

morphological changes in the connectivity of zonula occluden tight junction proteins and a decrease in TEER [171]. In our study of the brain endothelium, we went beyond qualitative observations of tight junction proteins and quantified the integrity of the barrier formed by the WT- and ccALD-iBMECs using TEER. By this metric, we found that the barrier integrity of the ccALD-iBMECs was decreased compared to WT-iBMECs. Musolino et al. also observed an increase in TGFβ1 expression connected to the mislocalization of claudin-5. Interestingly, our transcriptome analysis also indicated increased TGFβ1 activity in the ccALD-iBMECs that could be contributing to the decreased barrier function. With ccALD, in contrast to other demyelinating disorders such as multiple sclerosis, demyelination is thought to precede BBB breakdown. Thus, it is possible that an initial subtle loss in the ability of the BBB to restrict passive transport, as seen with the ccALD-iBMECs in our study, could cause immune cell infiltration and leakage that accelerates demyelination in a feedback loop that eventually results in complete BBB breakdown [60]. In this context, increased matrix metalloproteinases in the cerebral spinal fluid of ccALD patients could also be contributing to further breakdown of the BBB [75, 69]. Inherent decreased BBB integrity could also begin to explain why head trauma can initiate the onset of ccALD. The lack of genotype-phenotype correlation is not explained by our model; however, our finding of an inherent decrease in BMEC integrity in ccALD individuals could direct the search for additional environmental or genetic factors specific to the BBB that begin to explain why only a subset of individuals with an *ABCD1* mutation progress to ccALD.

While we did not observe by TEM the classic crystalline aggregates first observed in the adrenal cortex, testis, and white matter of ccALD patients, our finding of increased accumulation of non-pathological lipid droplets in ccALD-iBMECs is novel and warrants further investigation [172, 173, 174]. A study by Schluter et al. (2012) showed that VLCFAs can trigger insulin desensitization characterized by oxidative stress and alteration of adipocytokine signaling pathways and chronic inflammation, culminating in changes similar to metabolic syndrome [175].

Increased insulin-like growth factor receptor signaling in the ccALD-iBMECs indicated by our transcriptome analysis further hints at metabolic dysfunction as a factor contributing to the ccALD phenotype. Another study by van de Beek et al. (2017) showed that exposure of X-ALD fibroblasts to VLCFAs resulted in endoplasmic reticulum stress correlated with an increase in lipid droplet deposition [176]. Both studies suggest that VLCFA accumulation would precede lipid droplet accumulation. A key question that then arises is whether lipid droplet accumulation contributes to the decreased BBB integrity in the ccALD-iBMECs and whether targeting non-VLCFA lipids would have therapeutic relevance in that it could potentially rescue the decrease in BBB integrity of ccALD patients.

Using this system to model the BBB, we achieved physiological levels of TEER. One limitation of our study, however, is that we only investigated one cell type, BMECs. Adding other cell types involved in the neurovascular unit such as pericytes, astrocytes, and neurons to our model could further inform ccALD-specific defects of the BBB. Nevertheless, the differences we found modeling the BBB using iBMECs were significant. These differences included a decrease in barrier integrity as well as an increase in lipid accumulation. Both of these findings represent potential biomarkers for brain endothelium health of X-ALD patients and provide a new direction in the search for molecular markers that indicate ccALD onset. Combining the findings from the results of this study with antioxidant therapy currently in clinical trials could provide a much-needed alternative treatment for patients with AMN at risk of converting to ccALD.

With our BBB model, we investigated whether amphiphilic block copolymers can improve defects in barrier function as such polymers have been reported to be able to improve function of many cell and tissue types under various injuries. The application of the diblock copolymer $E_{182}P_{16}\underline{t}$ in addition to the widely used P188 was inspired by recent work within our group, which revealed $E_{182}P_{16}\underline{t}$ to be the most efficacious in stabilizing damaged myoblasts *in vitro* [85]. It is interesting and crucial to note that this enhanced efficacy of $E_{182}P_{16}\underline{t}$ compared to P188 is consistent with the results of Kim et al. although the cell types and form of damage

are vastly different [85]. This raises many questions as to how the polymers interact with the cell and what cellular responses this interaction promotes. At present, the fundamental mechanism of interaction between the PEO-PPO diblock and triblock copolymers and the plasma membrane is far from conclusive. Lee and co-workers speculate that poloxamers insert partially or fully into the membrane after initially adsorbing onto the lipid bilayer depending on the hydrophobicity of the polymer and the incubation time [177, 178]. Enhanced efficacy with the presence of an additional hydrophobic tert-butoxy end group on the PPO block provides evidence for the "anchor and chain" mechanism, which proposes that the additional hydrophobic unit at the end of the PPO block provides an anchor in the lipid bilayer, resulting in more efficacious stabilization of the block copolymer in the lipid membrane [101, 85]. The aforementioned studies focus on the polymer interaction with the plasma membrane for stabilization, but the results of our work showing the decrease in lipid accumulation in ccALD-iBMECs upon treatment with $E_{182}P_{16}\underline{t}$ suggest a more complex cellular response that has yet to be fully explored.

As the breadth of applications continues to expand, there is a pressing need to elucidate the amphiphilic block copolymer-cell interaction mechanism in order to translate this action into a therapeutic solution. To this end, the *in vitro* disease model presented in this work could provide a platform for studying the mechanism of PEO-PPO block copolymer mediated recovery of cellular function. Furthermore, while most researchers have focused solely on P188, there is potential for the design of the polymer to further improve efficacy in restoring function to damaged cells as demonstrated in this work. Elucidating the mechanism of BMEC interaction with the PEO-PPO diblock copolymer will not only engender insight as to how the polymer restores function of ccALD-iBMECs but may also provide a deeper knowledge as to how the BBBs of ccALD patients are damaged compared to healthy individuals.

At present, there is no suitable *in vivo* model for X-ALD [76]. Nevertheless,

as functional *in vivo* models are developed, the work presented here has the potential to be translated to *in vivo* studies. Treatment of ccALD-iBMECs with either P188 or $E_{182}P_{16}\underline{t}$ at the end of the differentiation protocol yielded a slight but non-significant increase in TEER. However, efficacy of polymer treatment at a later stage might be improved upon optimization of pharmacodynamics and pharmacokinetic variables. Furthermore, the superior efficacy of treatment with $E_{182}P_{16}\underline{t}$ when added earlier in the iBMEC differentiation process compared to at the end of the differentiation process suggests that the treatment could be applied at an early stage of BBB development to inhibit the onset and progression of ccALD.

Thus, clinical application might mean using the polymer as a preventative therapy, which requires pre-symptomatic diagnosis of X-ALD. Fortunately, high throughput screening of X-ALD is feasible and reliably identifies affected males [179]. Furthermore, in 2016 the US Department of Health and Human Services recommended that X-ALD be added to the recommended uniform screening panel for state newborn screening programs [180]. Testing of the 4 million infants born each year in the US is predicted to identify around 143 newborns with an *ABCD1* mutation. Early detection will lead to more timely intervention in the form of hematopoietic cell transplant (HCT), which is only advantageous in the early stages of the disease because cerebral inflammation can progress up to 18 months after transplant [57, 181]. Pioneering clinical trials involving the use of Lenti-D for autologous HCT are taking place at several centers around the US and promise to further reduce severe outcomes associated with allogeneic transplants (such as graft-versus-host disease) and to circumvent issues with finding HLA matched donors (currently, cord blood grafts are used when a suitable donor cannot be found) [182, 108, 183]. For those displaying neurological symptoms or MRI abnormalities indicating ccALD onset, the current standard of care for HCT involves fully myoablative chemotherapy, a highly toxic procedure. If a treatment that prevents the onset of ccALD were available, a newborn identified as having an *ABCD1* mutation given this treatment may never show symptoms of ccALD

conversion and would not need to undergo HCT. In this study, we have shown that amphiphilic block copolymers are one such treatment with the potential to prevent the onset of ccALD and reduce the number of patients needing to undergo HCT.

3.5 Conclusion

Modeling the BBB of ccALD patients using iPSC-derived BMECs indicates that ccALD patients form a less intact BBB. These results open the door for the discovery of brain endothelium-specific molecular markers indicative of the onset of ccALD and for the development of treatment strategies targeted at the brain endothelium that could reduce the number of X-ALD patients who progress to ccALD. One such treatment strategy that we have shown can rescue defective ccALD-iBMECs barrier integrity is PEO-PPO block copolymers. These results have therapeutic implications for preventing the onset of ccALD.

Chapter 4

Purification of Revertant Mosaic Fibroblasts from a patient with Recessive Dystrophic Epidermolysis Bullosa using Synthetic MicroRNA Switches

4.1 Hypothesis

Recessive dystrophic epidermolysis bullosa (RDEB) is a severe, lethal skin disorder characterized by chronic skin blistering and aberrant wound healing. This disorder is caused by loss-of-function mutations in the critical extracellular matrix protein type VII Collagen (C7). One of the limited treatment options available is the use of C7 expressing stem cells or differentiated skin cells to replace C7 at the dermal-epidermal junction and restore the overall integrity of the skin architecture. However, this typically requires the use of gene therapy or allogeneic cells, which can be costly and cause adverse reactions in the recipient. In rare cases, genetic reversion causes natural gene correction of the underlying mutation leading to a

population or patch of cells that is phenotypically distinct or mosaic. While these cells may be useful clinically, isolating and purifying them has proven difficult. I hypothesized that C7 producing and non-producing cells could be separated using microRNA (miR) switches, a method developed by Shinya Yamanaka [184].

4.2 Methods

Blistered and mosaic human dermal fibroblasts from a male RDEB patient were obtained by skin biopsy and maintained in DMEM GlutaMAX^{TM} (Thermo Fisher Scientific) containing 10% fetal bovine serum (MilliporeSigma), 1% Pen/Strep (Thermo Fisher Scientific), and 1% MEM NEAA (Thermo Fisher Scientific). For immunocytochemistry, fibroblasts were grown on chamber slides, fixed with 4% paraformaldehyde for 15 minutes, treated with 0.2% TritonX (Sigma) in PBS for 15 minutes, blocked in 2% BSA in PBS for 1 hour, and incubated at room temperature for 2 hours with primary antibody Collagen VII (1:2000; generously provided by Drs. David Woodley and Mei Chen). Corresponding secondary antibody donkey anti rabbit IgG Cy3 (1:500; Jackson Immunoresearch) was applied for 1 hour. Confocal images taken on an Olympus BX61 (Olympus Optical).

Total RNA was isolated using the miRpremier^{TM} microRNA Isolation Kit (MilliporeSigma). Small RNA libraries were constructed using the TruSeq Small RNA sample preparation kit (Illumina) with three biological replicates (different passages) of blistered and mosaic fibroblasts. From these libraries, single-end 50 bp reads were generated using an Illumina MiniSeq. Adaptor sequences were trimmed using Cutadapt [185]. Trimmed reads were aligned to the human genome GRCH38 using Bowtie2 [186]. The featureCounts function [187] of Subread was used to identify reads mapping to mature miRs as annotated in the hsa.gff3 file from mirBase [188]. Lowly expressed miRs not meeting the requirement of CPM $>$ 0.5 in at least two samples were removed from further analysis. edgeR (Robinson 2009) was used to identify miRs differentially expressed between blistered and mosaic fibroblasts.

miR switches were generated for the 10 differentially expressed miRs. miR switches are an *in vitro* transcribed GFP reporter with a a 5' RNA seed sequence complimentary to each identified miR. miR mimics (Thermo Fisher Scientific) were used to confirm change in fluorescence was specific to miR activity.

4.3 Results and Discussion

Immunofluorescence was used to determine the amount of C7 expression in dermal fibroblasts isolated from blistered and mosaic regions of a male RDEB patient's skin. Fibroblasts from the mosaic region exhibited greater amounts of C7 expression compared to those from the blistered region (Figure 4.1(a)). To identify miRs differentially expressed between blistered and mosaic fibroblasts, we performed small RNA-sequencing in triplicate on these fibroblast populations. While the number of features (miRs) identified for each sample varied, the distribution of raw read counts was similar after log transformation (see Supplementary Figures S15 and S16) and the blistered and mosaic samples separated well along the first dimension when visualized using an MDS plot (Figure 4.1(b)). Therefore, we performed statistical tests using edgeR and identified 10 miR candidates that were differentially expressed (FDR ≤ 0.05) with greater than twofold change between the blistered and mosaic samples (Figure 4.1(c) and Supplementary Table S1). We next performed qRT-PCR on the miRs with the lowest p-values, miR-#1 and miR-#2 to confirmed direction of expression (see Supplementary Figure S17).

When transfected into a mixed pool of C7 expressing and non-expressing fibroblasts, changes in fluorescence due to the activity of endogenous miRs on the miR switches should separate these two subpopulations. Flow cytometry results of untransfected cells, cells transfected with control GFP, mCherry, or both, and cells transfected with miR switches with or without miR mimics shown in Figure 4.2.

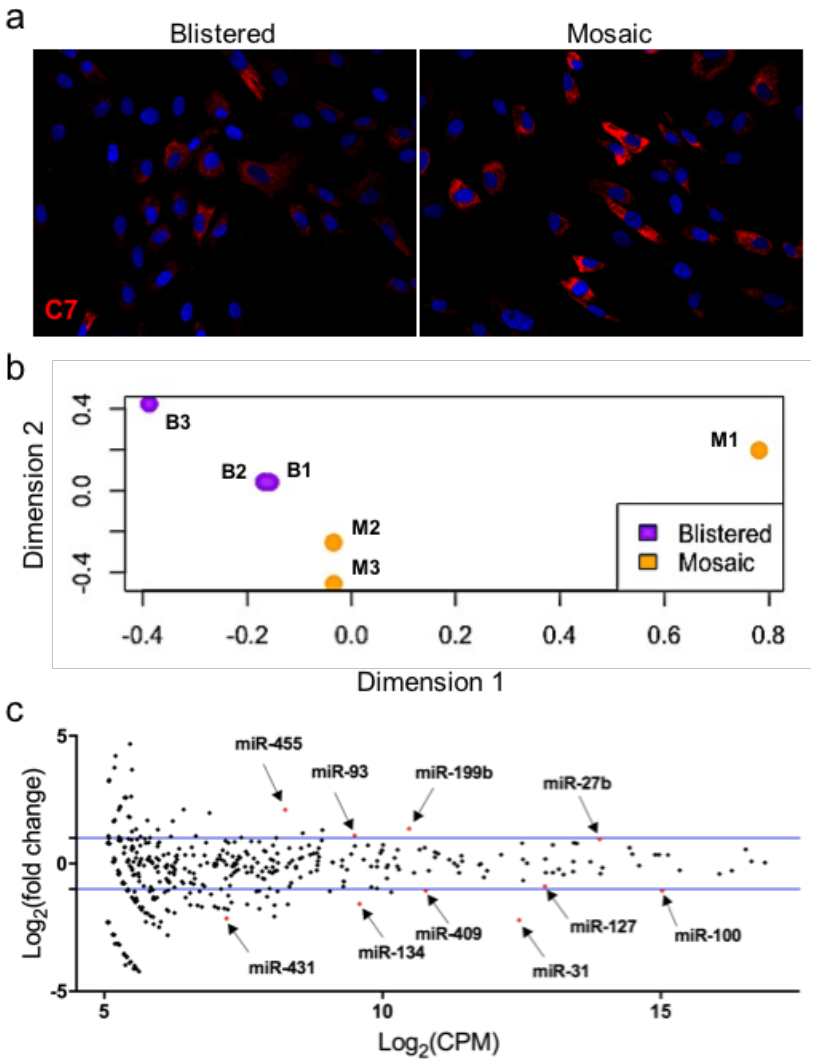


Figure 4.1: **Identification of differentially expressed miRs between blistered and mosaic fibroblast populations.** (a) Immunocytochemistry staining for C7 of blistered (left) and mosaic (right) fibroblasts. Nuclei counterstained with DAPI. Scale bar = 50 μm. (b) Multidimensional scaling (MDS) analysis of blistered and mosaic fibroblast populations following small RNA sequencing. Blistered samples shown in purple; mosaic samples shown in orange. (c) MA plot showing $\log_2$(fold change) versus $\log_2$(average CPM). Red dots indicate miRs found to be differentially expressed (FDR $\leq$ 0.05). Blues lines indicate 2X fold change cutoff.

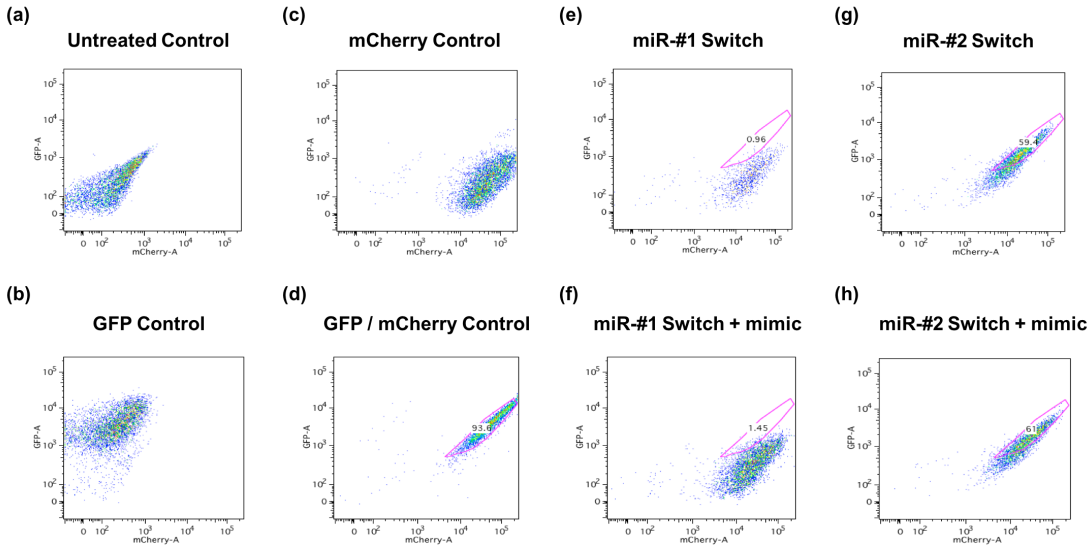


Figure 4.2: **Flow cytometry on fibroblasts transfected with miR switches.** (a) Untransfected control cells (b) cells transfected with naked GFP mRNA (c) cells transfected with naked mCherry mRNA (d) cells transfected with both GFP and mCherry control naked mRNA (e) cells transfected with mir-#1 switch (f) cells transfected with mir-#1 switch plus mir-#1 mimic (g) cells transfected with mir-#2 switch (h) cells transfected with mir-#2 switch plus mir-#2 mimic.

4.4 Conclusion and Future Directions

Based on preliminary experiments, further optimization of the protocol and miR switches is needed, including a dose-response experiment with the miR mimics. Future work includes flow sorting and Western blotting to confirm C7 expression in the mosaic fibroblast sub-populations followed by development of Puromycin miR switches to select on C7 expressing cells in culture. If successful, this technique will be useful in generating pure populations of mosaic cells that can then be used in future clinical applications such as expansion of the naturally-gene corrected cells for autologous hematopoietic cell transplant or skin grafts for areas of the body that fail to heal properly. This would make possible a non-palliative treatment option that is safer and less costly for patients suffering from this devastating disorder.

Chapter 5

A case study of RDEB mosaicism

5.1 Hypothesis

In rare cases, RDEB patients will have patches of skin that will appear completely normal. Patients report that these patches have never blistered. Examination at the molecular level shows functional C7 at these mosaic sites and revertant mosaicism at the DNA level. Revertant mosaicism is caused by post-zygotic autosomal mutations where one COL7A1 allele undergoes a back mutation or gene conversion event resulting in a patch of cells expressing functional C7 and reversion of the RDEB phenotype at this location. C7 is produced by epidermal keratinocytes and dermal fibroblasts. To date, only keratinocytes have been shown to be responsible for revertant mosaicism in RDEB, however, immunostaining data from our lab shows C7 expression by both keratinocytes and fibroblasts in one particular mosaic RDEB patient (Figure 5.1). I hypothesized that a revertant mosaic event had occurred that produced enough C7 to restore functionality in both the keratinocytes and fibroblasts.

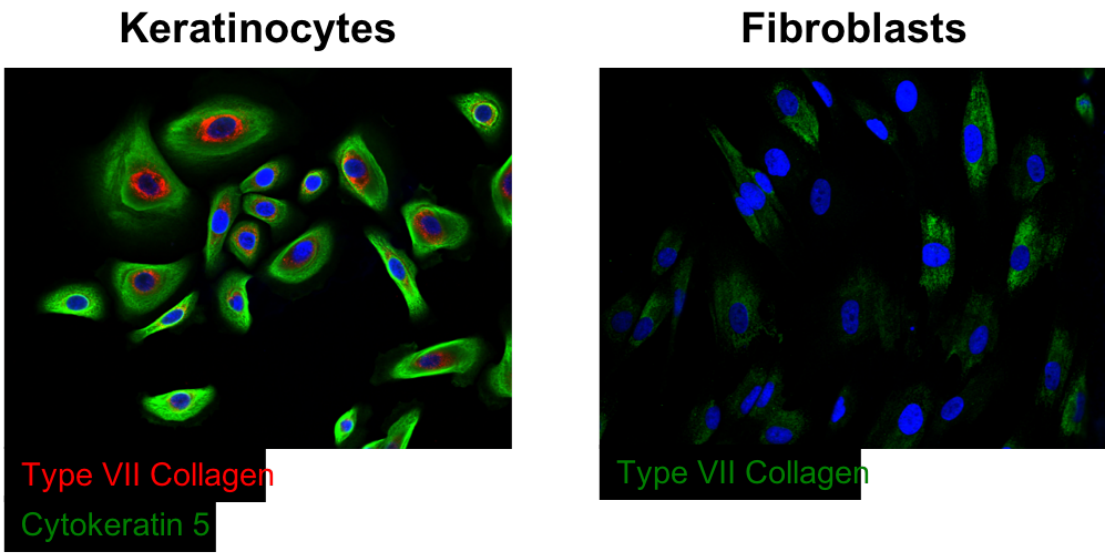


Figure 5.1: **Immunocytochemistry on mosaic fibroblasts and keratinocytes.** [Description of figure]

5.2 Methods

To determine which cell type is producing enough C7 to restore functionality I used Single-Molecule, Real-Time (SMRT) sequencing (PacBio sequencing) to query the full length of the C7 transcript from fibroblasts and keratinocytes taken from mosaic and blistered sites as well as a wild-type control. I then used the SMRT Portal provided by MSI to produce FASTQ files of the long-read sequencing data followed by mapping with BWA-MEM on MSI systems to produce BAM files for visualization.

5.3 Results

This particular patient is a compound heterozygote with a point mutation causing a premature termination codon (PTC) on one COL7A1 allele and a 16 bp deletion on the other allele predicted to cause a frameshift mutation and downstream PTC.

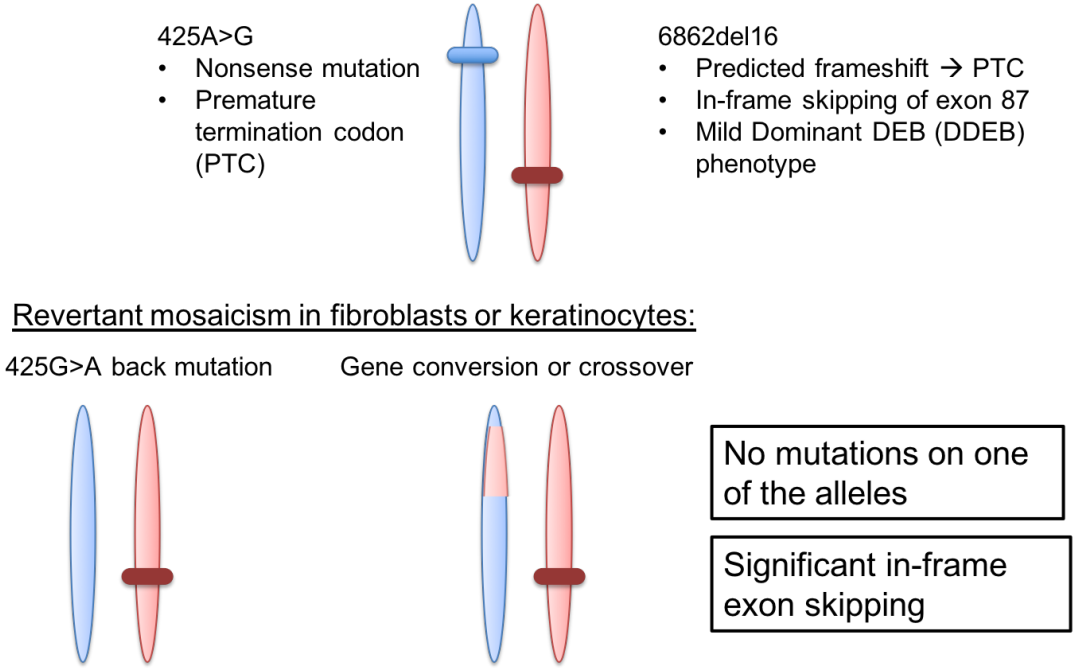


Figure 5.2: **Expected results after long-read sequencing.** [Description of figure]

Previous work has indicated that instead of a PTC, this 16 bp deletion causes in-frame exon skipping producing a truncated C7 transcript resulting in a milder dominant dystrophic epidermolysis bullosa (DDEB) phenotype. If a reversion event had occurred, I would expect to see a back mutation or gene conversion event evidenced by no mutations on one of the C7 alleles (Figure 5.2).

Full length transcripts were produced for all of the samples submitted (Figure 5.3). Read lengths produced shown in S18. Compared to the wild-type control, I saw no expression of the allele containing the point mutation indicating nonsense-mediated decay (Figure 5.4). For the other allele at the 16 bp deletion site, I saw large amounts of in-frame exon skipping in all of the patient samples but not in the wild-type control (Figure 5.5). 30% of normal C7 expression is enough to produce normal skin, however, it is not known how much truncated C7 protein is required. To confirm that no back mutations or gene conversions had taken place, I used DNA-sequencing (Table 5.1)

	6862del16	WT	425A>G	WT
Affected FB	4	2	0	0
Mosaic FB	3	2	0	2
Affected KT	4	3	0	2
Mosaic KT	1	1	0	0

Table 5.1: DNA-sequencing to confirm no back mutations or gene conversions have taken place.

Current work involves quantifying expression in this region to determine if levels of truncated C7 production differ between the keratinocytes and fibroblasts and the mosaic and blistered sites. I am doing this using short read sequencing targeted to C7. Since the C7 mRNA transcript is 9287 in length with 118 exons and a highly repetitive triple helical domain, a custom-designed Illumina panel with 68 probes was designed to quantify expression in this region. The short reads from this panel can be mapped to the long reads from the PacBio sequencing (i.e. the PacBio sequence will be used as the reference sequence).

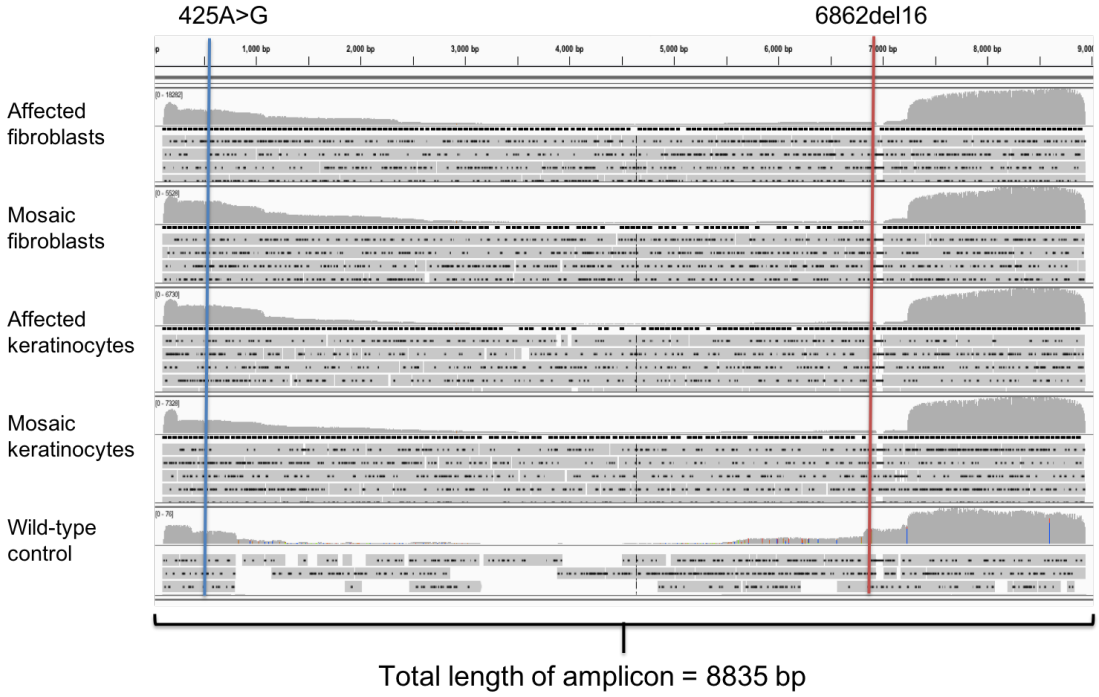


Figure 5.3: **Full length of C7 transcript.** [Description of figure]

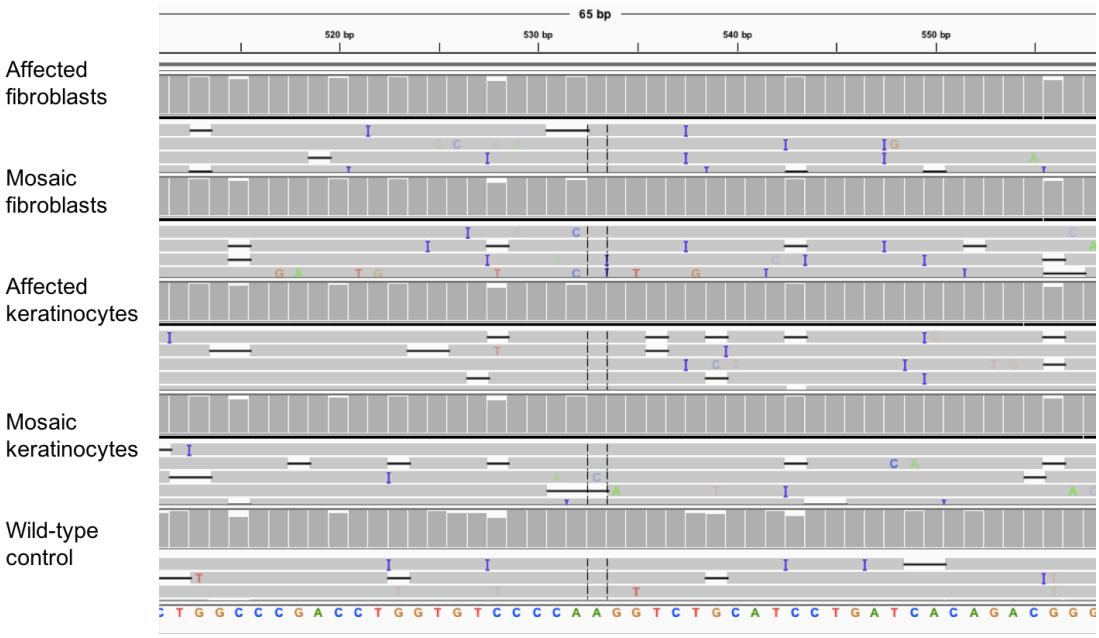


Figure 5.4: **Sequencing results across 425>G allele.** [Description of figure]

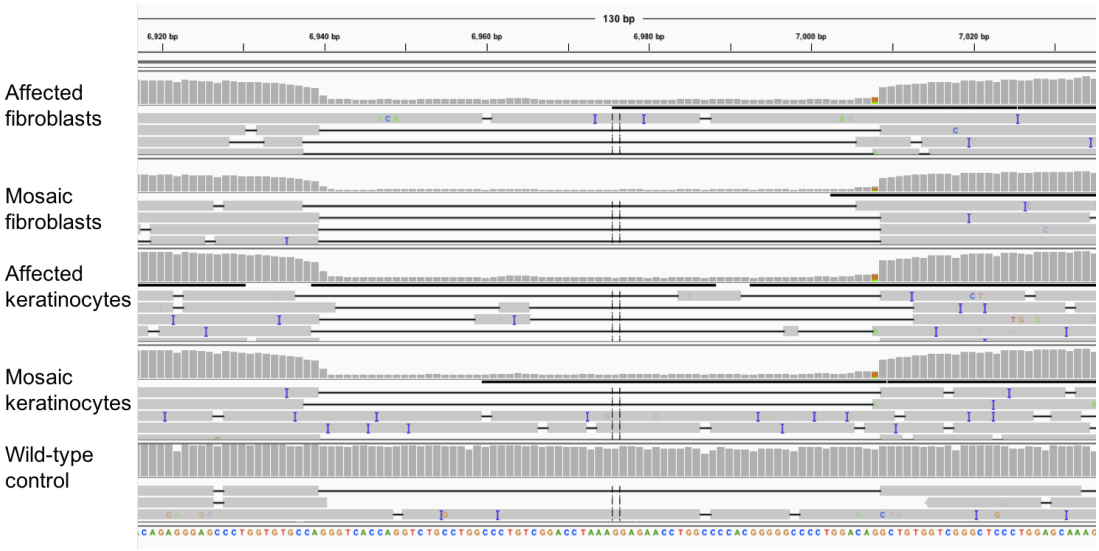


Figure 5.5: **Sequencing results across 6862del16 allele.** [Description of figure]

5.4 Conclusions and Future Directions

Using PacBio long-read sequencing, I was able to interrogate the entire length of the C7 transcript. Extremely low levels of the 425A>G mutation indicated nonsense-mediated decay of the maternal allele while significant in-frame exon skipping in both fibroblasts and keratinocytes at mosaic and affected sites was observed for the paternal allele. Current work includes quantifying expression over this region to determine if mosaic sites are reaching a functional threshold for C7 expression that is restoring skin integrity. This would be the first example of revertant mosaicism at the RNA level.

Future work includes DNA sequencing of the COL7A1 locus to confirm no back mutation or gene conversion has taken place and single-cell RNA-sequencing to determine if it is many cells expressing small amounts of C7 or a few cells expressing large amounts of C7 that is responsible for restoring functionality (Figure 5.6).

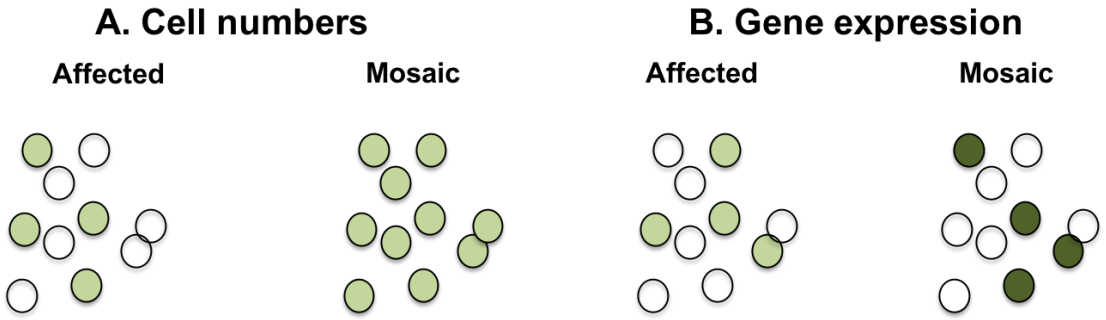


Figure 5.6: **Model of C7 expression in mosaic cells.** C7 expression due to in-frame skipping reaches functional threshold (hypomorphic expression) in mosaic sites. This could be the results of many cells expressing small amounts of C7 (A) or a few cells expressing large amounts of C7 (B).

This would give further insight into a novel mechanism of revertant mosaicism at the RNA level and would potentially be the first occurrence of mosaicism in fibroblasts rather than keratinocytes for this disease. Since fibroblasts are

easier to culture and expand than keratinocytes this has direct implications for therapeutic interventions. The life of this RDEB patient can be directly improved by expanding the C7 expressing cells in culture and creating 3D skin that can be grafted onto sites of frequent injury. Allogeneic skin grafts have been shown to increase C7 at the DEJ and provide relief to badly blistered areas. Using this patient's own cells mitigates the incidence of transplant-associated complications and would be a first example of "natural gene therapy.

Chapter 6

Conclusion

6.1 Summary

For this dissertation, I successfully identified *COL11A1, GREM1*, and *MFAP5* as markers of a subpopulation of fibroblasts in patients with recessive dystrophic epidermolysis bullosa (RDEB) using a novel, multitask clustering algorithm, scVDMC. Identification of these markers could lead to detection and early treatment of squamous cell carcinoma in individuals with RDEB. I modeled the blood-brain barrier (BBB) of patients with childhood cerebral adrenoleukodystrophy (ccALD) using directed differentiation of induced pluripotent stem cells (iPSCs) into brain microvascular endothelial cells (BMECs) and found that the iBMECs from patients with ccALD had significantly decreased TEER ($2592 \pm 110\ \Omega \cdot cm^2$) compared to WT controls ($5001 \pm 172\ \Omega \cdot cm^2$). I also found that they accumulate lipid droplets. Treatment with a PEO-PPO diblock copolymer reversed both of these defects. Improvements in barrier function produced by the PEO-PPO diblock copolymer has implications for translation into a treatment for preventing the onset of ccALD. Translating the results from this study has the potential to reduce the number of individuals who develop deadly and rapidly progressive ccALD.

The miR switches study demonstrates an applied technique for using next generation sequencing to identify markers of genetic heterogeneity. This technique

may be useful for generating pure populations of naturally gene corrected mosaic cells that can then be used in future clinical applications for RDEB patients such as expansion of the naturally-gene corrected cells for autologous hematopoietic cell transplant or skin grafts for areas of the body that fail to heal properly.

Going further, I also successfully used third generation sequencing technology, PacBio sequencing, to examine the underlying genetic heterogeneity in a case study of RDEB mosaicism. Quantifying these results could identity a new mechanism of revertant mosaicism.

6.2 The future of sequencing

In addition to long-read sequencing, Nanopore sequencing is another highly anticipated third-generation sequencing technology [4]. Oxford Nanopore Technologies (ONT) was the first company to market nanopore sequencers, the GridION and MinION [189]. The MinION especially has created much excitement as it is a USB-style device about the size of a mobile phone and has already been used to generate bacterial genome reference sequences [190] and sequence targeted amplicons [191]. The fast run times and compact nature of the device present an opportunity to democratize sequencing, removing it from the domain of core services available only at select research institutions and placing the technology in the hands of many separate sources [4]. This is well exemplified in the recent work of scientists who used the MinION in the field to sequence the Ebola virus in Guinea two days after sample collection [192, 193].

Even core genomic facilities are moving towards purchasing many smaller genomics instruments rather than a few expensive, higher capacity instruments. Researchers would not need to submit samples to be processed by the facility but instead would use the instruments via direct access at kiosks within the core facility or by renting them for use in their own lab. This initiative to democratize genomics will lead to exponentially greater amounts of genetic information being readily and affordably available and will lead to further extension of information

across cell types, individuals, and populations, which in turn will enable precision medicine to become the standard of care.

Recently, two new droplet-based RNA-seq technologies, Drop-seq and inDrops (indexing droplets) [128, 194], can capture and sequence thousands of single-cells quickly and cost effectively using nanoliter-scale aqueous droplets containing a single-cell, barcoded and UMI-labeled primers, and reaction buffer. STAMPs (Single-cell Transcriptomes Attached to Microparticles) are PCR amplified for sequencing in Drop-seq, while Cell-seq is used by inDrops for sequencing. The latest commercial platform, the Chromium^{TM} System from 10X Genomics, integrates the Gemcode platform, which separates long pieces of DNA into droplets to create barcoded sequencing libraries [99].

As next generation and single-cell sequencing technology becomes more accessible and affordable to more researchers, the amount of information that is available will only continue to increase. Meta-analysis of this information will allow systems biology to build increasingly more complex models of genetic, transcriptomic, and proteomic interactions and to integrate information across not only cells, but across patients and populations as well. The technology to decipher the relative position of single-cells is much needed. When this arrives, it is possible that testing of therapeutic interventions such as drugs, small molecules, or genetic edits may not be done on laboratory animals, but on *in silico* human bodies that use machine learning techniques to learn from and build on the plethora of information that the sequencing era has left us, ushering in the next generation of precision medicine.

References

[1] D. L. Nelson. Turning the corner from observation to intervention in human genetics. *J Genet Genomics*, 45(2):57–59, Feb 2018.

[2] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[3] B. Maddox and R. Franklin. The double helix and the 'wronged heroine'. *Nature*, 421(6921):407–408, Jan 2003.

[4] J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, Jan 2016.

[5] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, 6(7):2601–2610, Jun 1979.

[6] S. Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.*, 9(13):3015–3027, Jul 1981.

[7] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986.

[8] No authors listed. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[9] M. Ronaghi, B. Pettersson, M. Uhlen, and P. Nyren. PCR-introduced loop structure as primer in DNA sequencing. *BioTechniques*, 25(5):876–878, Nov 1998.

[10] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34(3):e22, Feb 2006.

[11] G. Turcatti, A. Romieu, M. Fedurco, and A. P. Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.*, 36(4):e25, Mar 2008.

[12] L. D. Stein. The case for cloud computing in genome informatics. *Genome Biol.*, 11(5):207, 2010.

[13] K. M. Boycott, A. Rath, J. X. Chong, T. Hartley, F. S. Alkuraya, G. Baynam, A. J. Brookes, M. Brudno, A. Carracedo, J. T. den Dunnen, S. O. M. Dyke, X. Estivill, J. Goldblatt, C. Gonthier, S. C. Groft, I. Gut, A. Hamosh, P. Hieter, S. Hohn, M. E. Hurles, P. Kaufmann, B. M. Knoppers, J. P. Krischer, M. Macek, G. Matthijs, A. Olry, S. Parker, J. Paschall, A. A. Philippakis, H. L. Rehm, P. N. Robinson, P. C. Sham, R. Stefanov, D. Taruscio, D. Unni, M. R. Vanstone, F. Zhang, H. Brunner, M. J. Bamshad, and H. Lochmuller. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am. J. Hum. Genet.*, 100(5):695–705, May 2017.

[14] M. Titeux, A. Izmiryan, and A. Hovnanian. The Molecular Revolution in Cutaneous Biology: Emerging Landscape in Genomic Dermatology: New Mechanistic Ideas, Gene Editing, and Therapeutic Breakthroughs. *J. Invest. Dermatol.*, 137(5):e123–e129, May 2017.

[15] F. Niu, D. C. Wang, J. Lu, W. Wu, and X. Wang. Potentials of single-cell biology in identification and validation of disease biomarkers. *J. Cell. Mol. Med.*, 20(9):1789–1795, 09 2016.

[16] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[17] United States 107th Congress. H.R. 4013 - Rare Diseases Act of 2002. Accessed: 2018-04-12.

[18] Council of European Union. Council regulation (EU) No 141/2000, 1999. Accessed: 2018-04-12.

[19] V. Wally, S. Kitzmueller, F. Lagler, A. Moder, W. Hitzl, M. Wolkersdorfer, P. Hofbauer, T. K. Felder, M. Dornauer, A. Diem, N. Eiler, and J. W. Bauer. Topical diacerein for epidermolysis bullosa: a randomized controlled pilot study. *Orphanet J Rare Dis*, 8:69, May 2013.

[20] A. Hovnanian, P. Duquesnoy, C. Blanchet-Bardon, R. G. Knowlton, S. Amselem, M. Lathrop, L. Dubertret, J. Uitto, and M. Goossens. Genetic linkage of recessive dystrophic epidermolysis bullosa to the type VII collagen gene. *J. Clin. Invest.*, 90(3):1032–1036, Sep 1992.

[21] D. R. Keene, L. Y. Sakai, G. P. Lunstrum, N. P. Morris, and R. E. Burgeson. Type VII collagen forms an extended network of anchoring fibrils. *J. Cell Biol.*, 104(3):611–621, Mar 1987.

[22] B. R. Webber and J. Tolar. From marrow to matrix: novel gene and cell therapies for epidermolysis bullosa. *Mol. Ther.*, 23(6):987–992, Jun 2015.

[23] H. M. Horn and M. J. Tidman. Quality of life in epidermolysis bullosa. *Clin. Exp. Dermatol.*, 27(8):707–710, Nov 2002.

[24] L. Bruckner-Tuderman. Dystrophic epidermolysis bullosa: pathogenesis and clinical features. *Dermatol Clin*, 28(1):107–114, Jan 2010.

[25] R. B. Petersen, C. T. Bonde, and G. Schmidt. Generalized dystrophic epidermolysis bullosa and development of squamous cell carcinoma. *Ugeskr. Laeg.*, 172(47):3267–3268, Nov 2010.

[26] J. E. Wagner, A. Ishida-Yamamoto, J. A. McGrath, M. Hordinsky, D. R. Keene, D. T. Woodley, M. Chen, M. J. Riddle, M. J. Osborn, T. Lund, M. Dolan, B. R. Blazar, and J. Tolar. Bone marrow transplantation for recessive dystrophic epidermolysis bullosa. *N. Engl. J. Med.*, 363(7):629–639, Aug 2010.

[27] M. J. Vanden Oever and J. Tolar. Advances in understanding and treating dystrophic epidermolysis bullosa. *F1000Prime Rep*, 6:35, 2014.

[28] F. Mavilio, G. Pellegrini, S. Ferrari, F. Di Nunzio, E. Di Iorio, A. Recchia, G. Maruggi, G. Ferrari, E. Provasi, C. Bonini, S. Capurro, A. Conti, C. Magnoni, A. Giannetti, and M. De Luca. Correction of junctional epidermolysis bullosa by transplantation of genetically modified epidermal stem cells. *Nat. Med.*, 12(12):1397–1402, Dec 2006.

[29] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 1993 Jan 01 - . Identifier NCT01263379. Gene Transfer for Recessive Dystrophic Epidermolysis Bullosa.

[30] M. Goto, D. Sawamura, W. Nishie, K. Sakai, J. R. McMillan, M. Akiyama, and H. Shimizu. Targeted skipping of a single exon harboring a premature termination codon mutation: implications and potential for gene correction therapy for selective dystrophic epidermolysis bullosa patients. *J. Invest. Dermatol.*, 126(12):2614–2620, Dec 2006.

[31] S. Turczynski, M. Titeux, N. Pironon, H. I. Cohn, D. F. Murrell, and A. Hovnanian. Marked intrafamilial phenotypic heterogeneity in dystrophic epidermolysis bullosa caused by inheritance of a mild dominant glycine substitution and a novel deep intronic recessive COL7A1 mutation. *Br. J. Dermatol.*, 174(5):1122–1125, May 2016.

[32] L. Bezman, A. B. Moser, G. V. Raymond, P. Rinaldo, P. A. Watkins, K. D. Smith, N. E. Kass, and H. W. Moser. Adrenoleukodystrophy: incidence, new mutation rate, and results of extended family screening. *Ann. Neurol.*, 49(4):512–517, Apr 2001.

[33] A. M. Douar, J. Mosser, C. O. Sarde, J. Lopez, J. L. Mandel, and P. Aubourg. X-linked adrenoleukodystrophy gene: identification of a candidate gene by positional cloning. *Biomed. Pharmacother.*, 48(5-6):215–218, 1994.

[34] P. Aubourg, J. Mosser, A. M. Douar, C. O. Sarde, J. Lopez, and J. L. Mandel. Adrenoleukodystrophy gene: unexpected homology to a protein involved in peroxisome biogenesis. *Biochimie*, 75(3-4):293–302, 1993.

[35] J. Mosser, A. M. Douar, C. O. Sarde, P. Kioschis, R. Feil, H. Moser, A. M. Poustka, J. L. Mandel, and P. Aubourg. Putative X-linked adrenoleukodystrophy gene shares unexpected homology with ABC transporters. *Nature*, 361(6414):726–730, Feb 1993.

[36] F. Kok, S. Neumann, C. O. Sarde, S. Zheng, K. H. Wu, H. M. Wei, J. Bergin, P. A. Watkins, S. Gould, and G. Sack. Mutational analysis of patients with X-linked adrenoleukodystrophy. *Hum. Mutat.*, 6(2):104–115, 1995.

[37] G. C. Korenke, H. J. Christen, B. Kruse, D. H. Hunneman, and F. Hanefeld. Progression of X-linked adrenoleukodystrophy under interferon-beta therapy. *J. Inherit. Metab. Dis.*, 20(1):59–66, Mar 1997.

[38] J. J. Martin, B. Dompas, C. Ceuterick, and K. Jacobs. Adrenomyeloneuropathy and adrenoleukodystrophy in two brothers. *Eur. Neurol.*, 19(5):281–287, 1980.

[39] G. Sobue, I. Ueno-Natsukari, H. Okamoto, T. A. Connell, I. Aizawa, K. Mizoguchi, M. Honma, G. Ishikawa, T. Mitsuma, and N. Natsukari. Phenotypic heterogeneity of an adult form of adrenoleukodystrophy in monozygotic twins. *Ann. Neurol.*, 36(6):912–915, Dec 1994.

[40] H. W. Moser. Adrenoleukodystrophy. *Curr. Opin. Neurol.*, 8(3):221–226, Jun 1995.

[41] H. Moser, P. Dubey, and A. Fatemi. Progress in X-linked adrenoleukodystrophy. *Curr. Opin. Neurol.*, 17(3):263–269, Jun 2004.

[42] H. W. Moser, G. V. Raymond, and P. Dubey. Adrenoleukodystrophy: new approaches to a neurodegenerative disease. *JAMA*, 294(24):3131–3134, Dec 2005.

[43] H. W. Moser, A. Mahmood, and G. V. Raymond. X-linked adrenoleukodystrophy. *Nat Clin Pract Neurol*, 3(3):140–151, Mar 2007.

[44] M. L. Chu, D. A. Sala, and H. L. Weiner. Intrathecal baclofen in X-linked adrenoleukodystrophy. *Pediatr. Neurol.*, 24(2):156–158, Feb 2001.

[45] E. Shapiro, W. Krivit, L. Lockman, I. Jambaque, C. Peters, M. Cowan, R. Harris, S. Blanche, P. Bordigoni, D. Loes, R. Ziegler, M. Crittenden, D. Ris, B. Berg, C. Cox, H. Moser, A. Fischer, and P. Aubourg. Long-term effect of bone-marrow transplantation for childhood-onset cerebral X-linked adrenoleukodystrophy. *Lancet*, 356(9231):713–718, Aug 2000.

[46] V A Drover. Adrenoleukodystrophy: Recent advances in treatment and disease etiology. *Future Lipidology*, 4(2):205–213, 2009.

[47] B. Wilken, P. Dechent, K. Brockmann, J. Finsterbusch, M. Baumann, W. Ebell, G. C. Korenke, P. J. Pouwels, F. A. Hanefeld, and J. Frahm. Quantitative proton magnetic resonance spectroscopy of children with adrenoleukodystrophy before and after hematopoietic stem cell transplantation. *Neuropediatrics*, 34(5):237–246, Jun 2003.

[48] F. Eichler and K. Van Haren. Immune response in leukodystrophies. *Pediatr. Neurol.*, 37(4):235–244, Oct 2007.

[49] F. S. Eichler, R. Itoh, P. B. Barker, S. Mori, E. S. Garrett, P. C. van Zijl, H. W. Moser, G. V. Raymond, and E. R. Melhem. Proton MR spectroscopic and diffusion tensor brain MR imaging in X-linked adrenoleukodystrophy: initial experience. *Radiology*, 225(1):245–252, Oct 2002.

[50] W. P. Miller, L. F. Mantovani, J. Muzic, J. B. Rykken, R. S. Gawande, T. C. Lund, R. M. Shanley, G. V. Raymond, P. J. Orchard, and D. R. Nascene. Intensity of MRI Gadolinium Enhancement in Cerebral Adrenoleukodystrophy: A Biomarker for Inflammation and Predictor of Outcome following Transplantation in Higher Risk Patients. *AJNR Am J Neuroradiol*, 37(2):367–372, Feb 2016.

[51] P. L. Musolino, O. Rapalino, P. Caruso, V. S. Caviness, and F. S. Eichler. Hypoperfusion predicts lesion progression in cerebral X-linked adrenoleukodystrophy. *Brain*, 135(Pt 9):2676–2683, Sep 2012.

[52] E. Ratai, T. Kok, C. Wiggins, G. Wiggins, E. Grant, B. Gagoski, G. O'Neill, E. Adalsteinsson, and F. Eichler. Seven-Tesla proton magnetic resonance spectroscopic imaging in adult X-linked adrenoleukodystrophy. *Arch. Neurol.*, 65(11):1488–1494, Nov 2008.

[53] A. Budhram and S. K. Pandey. Activation of Cerebral X-linked Adrenoleukodystrophy After Head Trauma. *Can J Neurol Sci*, 44(5):597–598, 09 2017.

[54] G. V. Raymond, R. Seidman, T. S. Monteith, E. Kolodny, S. Sathe, A. Mahmood, and J. M. Powers. Head trauma can initiate the onset of adrenoleukodystrophy. *J. Neurol. Sci.*, 290(1-2):70–74, Mar 2010.

[55] I. A. Wilkinson, I. J. Hopkins, and A. C. Pollard. Can head injury influence the site of demyelination in adrenoleukodystrophy? *Dev Med Child Neurol*, 29(6):797–800, Dec 1987.

[56] W. P. Miller, S. M. Rothman, D. Nascene, T. Kivisto, T. E. DeFor, R. S. Ziegler, J. Eisengart, K. Leiser, G. Raymond, T. C. Lund, J. Tolar, and P. J. Orchard. Outcomes after allogeneic hematopoietic cell transplantation for childhood cerebral adrenoleukodystrophy: the largest single-institution cohort report. *Blood*, 118(7):1971–1978, Aug 2011.

[57] B. R. Turk, A. B. Moser, and A. Fatemi. Therapeutic strategies in adrenoleukodystrophy. *Wien Med Wochenschr*, 167(9-10):219–226, Jun 2017.

[58] I. Ferrer, P. Aubourg, and A. Pujol. General aspects and neuropathology of X-linked adrenoleukodystrophy. *Brain Pathol.*, 20(4):817–830, Jul 2010.

[59] C. Wiesinger, F. S. Eichler, and J. Berger. The genetic landscape of X-linked adrenoleukodystrophy: inheritance, mutations, modifier genes, and diagnosis. *Appl Clin Genet*, 8:109–121, 2015.

[60] P. Aubourg. Cerebral adrenoleukodystrophy: a demyelinating disease that leaves the door wide open. *Brain*, 138(Pt 11):3133–3136, Nov 2015.

[61] S. Forss-Petter, H. Werner, J. Berger, H. Lassmann, B. Molzer, M. H. Schwab, H. Bernheimer, F. Zimmermann, and K. A. Nave. Targeted inactivation of the X-linked adrenoleukodystrophy gene in mice. *J. Neurosci. Res.*, 50(5):829–843, Dec 1997.

[62] S. Fourcade, I. Ferrer, and A. Pujol. Oxidative stress, mitochondrial and proteostasis malfunction in adrenoleukodystrophy: A paradigm for axonal degeneration. *Free Radic. Biol. Med.*, 88(Pt A):18–29, Nov 2015.

[63] B. M. van Geel, J. Assies, E. B. Haverkort, J. H. Koelman, B. Verbeeten, R. J. Wanders, and P. G. Barth. Progression of abnormalities in adrenomyeloneuropathy and neurologically asymptomatic X-linked adrenoleukodystrophy despite treatment with "Lorenzo's oil". *J. Neurol. Neurosurg. Psychiatry*, 67(3):290–299, Sep 1999.

[64] G. A. Horvath, F. Eichler, K. Poskitt, and S. Stockler-Ipsiroglu. Failure of repeated cyclophosphamide pulse therapy in childhood cerebral X-linked adrenoleukodystrophy. *Neuropediatrics*, 43(1):48–52, Feb 2012.

[65] M. Deon, A. Sitta, A. G. Barschak, D. M. Coelho, M. Pigatto, G. O. Schmitt, L. B. Jardim, R. Giugliani, M. Wajner, and C. R. Vargas. Induction of lipid peroxidation and decrease of antioxidant defenses in symptomatic and asymptomatic patients with X-linked adrenoleukodystrophy. *Int. J. Dev. Neurosci.*, 25(7):441–444, Nov 2007.

[66] S. Fourcade, J. Lopez-Erauskin, J. Galino, C. Duval, A. Naudi, M. Jove, S. Kemp, F. Villarroya, I. Ferrer, R. Pamplona, M. Portero-Otin, and A. Pujol. Early oxidative damage underlying neurodegeneration in X-adrenoleukodystrophy. *Hum. Mol. Genet.*, 17(12):1762–1773, Jun 2008.

[67] D. P. Marchetti, B. Donida, H. T. da Rosa, P. R. Manini, D. J. Moura, J. Saffi, M. Deon, C. P. Mescka, D. M. Coelho, L. B. Jardim, and C. R. Vargas. Protective effect of antioxidants on DNA damage in leukocytes from X-linked adrenoleukodystrophy patients. *Int. J. Dev. Neurosci.*, 43:8–15, Jun 2015.

[68] J. Singh and S. Giri. Loss of AMP-activated protein kinase in X-linked adrenoleukodystrophy patient-derived fibroblasts and lymphocytes. *Biochem. Biophys. Res. Commun.*, 445(1):126–131, Feb 2014.

[69] K. A. Thibert, G. V. Raymond, D. R. Nascene, W. P. Miller, J. Tolar, P. J. Orchard, and T. C. Lund. Cerebrospinal fluid matrix metalloproteinases are elevated in cerebral adrenoleukodystrophy and correlate with MRI severity and neurologic dysfunction. *PLoS ONE*, 7(11):e50430, 2012.

[70] J. Tolar, P. J. Orchard, K. J. Bjoraker, R. S. Ziegler, E. G. Shapiro, and L. Charnas. N-acetyl-L-cysteine improves outcome of advanced cerebral adrenoleukodystrophy. *Bone Marrow Transplant.*, 39(4):211–215, Feb 2007.

[71] F. J. Rockenbach, M. Deon, D. P. Marchese, V. Manfredini, C. Mescka, G. S. Ribas, C. T. Habekost, C. G. Castro, L. B. Jardim, and C. R. Vargas. The effect of bone marrow transplantation on oxidative stress in X-linked adrenoleukodystrophy. *Mol. Genet. Metab.*, 106(2):231–236, Jun 2012.

[72] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 1993 Jan 01 - . Identifier NCT01495260. A Clinical Trial for AMN: Validation of Biomarkers of Oxidative Stress, Efficacy and Safety of a Mixture of Antioxidants.

[73] J. Berger, S. Forss-Petter, and F. S. Eichler. Pathophysiology of X-linked adrenoleukodystrophy. *Biochimie*, 98:135–142, Mar 2014.

[74] M. Ito, B. M. Blumberg, D. J. Mock, A. D. Goodman, A. B. Moser, H. W. Moser, K. D. Smith, and J. M. Powers. Potential environmental and host participants in the early white matter lesion of adreno-leukodystrophy: morphologic evidence for CD8 cytotoxic T cells, cytolysis of oligodendrocytes, and CD1-mediated lipid antigen presentation. *J. Neuropathol. Exp. Neurol.*, 60(10):1004–1019, Oct 2001.

[75] P. L. Musolino, Y. Gong, J. M. Snyder, S. Jimenez, J. Lok, E. H. Lo, A. B. Moser, E. F. Grabowski, M. P. Frosch, and F. S. Eichler. Brain endothelial dysfunction in cerebral adrenoleukodystrophy. *Brain*, 138(Pt 11):3206–3220, Nov 2015.

[76] J. F. Lu, A. M. Lawler, P. A. Watkins, J. M. Powers, A. B. Moser, H. W. Moser, and K. D. Smith. A mouse model for X-linked adrenoleukodystrophy. *Proc. Natl. Acad. Sci. U.S.A.*, 94(17):9366–9371, Aug 1997.

[77] M. J. Bernas, F. L. Cardoso, S. K. Daley, M. E. Weinand, A. R. Campos, A. J. Ferreira, J. B. Hoying, M. H. Witte, D. Brites, Y. Persidsky, S. H. Ramirez, and M. A. Brito. Establishment of primary cultures of human brain microvascular endothelial cells to provide an in vitro cellular model of the blood-brain barrier. *Nat Protoc*, 5(7):1265–1272, Jul 2010.

[78] H. C. Helms, N. J. Abbott, M. Burek, R. Cecchelli, P. O. Couraud, M. A. Deli, C. Forster, H. J. Galla, I. A. Romero, E. V. Shusta, M. J. Stebbins, E. Vandenhaute, B. Weksler, and B. Brodin. In vitro models of the blood-brain barrier: An overview of commonly used brain endothelial cell culture models and guidelines for their use. *J. Cereb. Blood Flow Metab.*, 36(5):862–890, May 2016.

[79] B. B. Weksler, E. A. Subileau, N. Perriere, P. Charneau, K. Holloway, M. Leveque, H. Tricoire-Leignel, A. Nicotra, S. Bourdoulous, P. Turowski, D. K. Male, F. Roux, J. Greenwood, I. A. Romero, and P. O. Couraud. Blood-brain barrier-specific properties of a human adult brain endothelial cell line. *FASEB J.*, 19(13):1872–1874, Nov 2005.

[80] E. S. Lippmann, A. Al-Ahmad, S. M. Azarin, S. P. Palecek, and E. V. Shusta. A retinoic acid-enhanced, multicellular human blood-brain barrier model derived from stem cell sources. *Sci Rep*, 4:4160, Feb 2014.

[81] E. S. Lippmann, S. M. Azarin, J. E. Kay, R. A. Nessler, H. K. Wilson, A. Al-Ahmad, S. P. Palecek, and E. V. Shusta. Derivation of blood-brain barrier endothelial cells from human pluripotent stem cells. *Nat. Biotechnol.*, 30(8):783–791, Aug 2012.

[82] M. J. Stebbins, H. K. Wilson, S. G. Canfield, T. Qian, S. P. Palecek, and E. V. Shusta. Differentiation and characterization of human pluripotent stem cell-derived brain microvascular endothelial cells. *Methods*, 101:93–102, 05 2016.

[83] H. K. Wilson, S. G. Canfield, M. K. Hjortness, S. P. Palecek, and E. V. Shusta. Exploring the effects of cell seeding density on the differentiation of human pluripotent stem cells to brain microvascular endothelial cells. *Fluids Barriers CNS*, 12:13, May 2015.

[84] R. G. Lim, C. Quan, A. M. Reyes-Ortiz, S. E. Lutz, A. J. Kedaigle, T. A. Gipson, J. Wu, G. D. Vatine, J. Stocksdale, M. S. Casale, C. N. Svendsen, E. Fraenkel, D. E. Housman, D. Agalliu, and L. M. Thompson. Huntington's Disease iPSC-Derived Brain Microvascular Endothelial Cells Reveal WNT-Mediated Angiogenic and Blood-Brain Barrier Deficits. *Cell Rep*, 19(7):1365–1377, 05 2017.

[85] B. J. Kim, O. B. Bee, M. A. McDonagh, M. J. Stebbins, S. P. Palecek, K. S. Doran, and E. V. Shusta. Modeling Group B Streptococcus and Blood-Brain Barrier Interaction by Using Induced Pluripotent Stem Cell-Derived Brain Endothelial Cells. *mSphere*, 2(6), 2017.

[86] I. R. Schmolka. Physical basis for poloxamer interactions. *Ann. N. Y. Acad. Sci.*, 720:92–97, May 1994.

[87] Irving R. Schmolka. A review of block polymer surfactants. *Journal of the American Oil Chemists' Society*, 54(3):110–116, Mar 1977.

[88] J. G. Moloughney and N. Weisleder. Poloxamer 188 (p188) as a membrane resealing reagent in biomedical applications. *Recent Pat Biotechnol*, 6(3):200–211, Dec 2012.

[89] R. C. Lee, L. P. River, F. S. Pan, L. Ji, and R. L. Wollmann. Surfactant-induced sealing of electropermeabilized skeletal muscle membranes in vivo. *Proc. Natl. Acad. Sci. U.S.A.*, 89(10):4524–4528, May 1992.

[90] B. Greenebaum, K. Blossfield, J. Hannig, C. S. Carrillo, M. A. Beckett, R. R. Weichselbaum, and R. C. Lee. Poloxamer 188 prevents acute necrosis of adult skeletal muscle cells following high-dose irradiation. *Burns*, 30(6):539–547, Sep 2004.

[91] S. W. Wong, Y. Yao, Y. Hong, Z. Ma, S. H. Kok, S. Sun, M. Cho, K. K. Lee, and A. F. Mak. Preventive Effects of Poloxamer 188 on Muscle Cell Damage Mechanics Under Oxidative Stress. *Ann Biomed Eng*, 45(4):1083–1092, 04 2017.

[92] S. Yasuda, D. Townsend, D. E. Michele, E. G. Favre, S. M. Day, and J. M. Metzger. Dystrophic heart failure blocked by membrane sealant poloxamer. *Nature*, 436(7053):1025–1029, Aug 2005.

[93] E. M. Houang, K. J. Haman, A. Filareto, R. C. Perlingeiro, F. S. Bates, D. A. Lowe, and J. M. Metzger. Membrane-stabilizing copolymers confer marked protection to dystrophic skeletal muscle in vivo. *Mol Ther Methods Clin Dev*, 2:15042, 2015.

[94] H. Bao, X. Yang, Y. Zhuang, Y. Huang, T. Wang, M. Zhang, D. Dai, S. Wang, H. Xiao, G. Huang, J. Kuai, and L. Tao. The effects of poloxamer 188 on the autophagy induced by traumatic brain injury. *Neurosci. Lett.*, 634:7–12, Nov 2016.

[95] G. Serbest, J. Horwitz, and K. Barbee. The effect of poloxamer-188 on neuronal cell recovery from mechanical injury. *J. Neurotrauma*, 22(1):119–132, Jan 2005.

[96] C. L. Luo, X. P. Chen, L. L. Li, Q. Q. Li, B. X. Li, A. M. Xue, H. F. Xu, D. K. Dai, Y. W. Shen, L. Y. Tao, and Z. Q. Zhao. Poloxamer 188 attenuates in vitro traumatic brain injury-induced mitochondrial and lysosomal membrane permeabilization damage in cultured primary neurons. *J. Neurotrauma*, 30(7):597–607, Apr 2013.

[97] H. J. Bao, T. Wang, M. Y. Zhang, R. Liu, D. K. Dai, Y. Q. Wang, L. Wang, L. Zhang, Y. Z. Gao, Z. H. Qin, X. P. Chen, and L. Y. Tao. Poloxamer-188 attenuates TBI-induced blood-brain barrier damage leading to decreased brain edema and reduced cellular death. *Neurochem. Res.*, 37(12):2856–2867, Dec 2012.

[98] J. H. Gu, J. B. Ge, M. Li, H. D. Xu, F. Wu, and Z. H. Qin. Poloxamer 188 protects neurons against ischemia/reperfusion injury through preserving integrity of cell membranes and blood brain barrier. *PLoS ONE*, 8(4):e61641, 2013.

[99] T. Wang, X. Chen, Z. Wang, M. Zhang, H. Meng, Y. Gao, B. Luo, L. Tao, and Y. Chen. Poloxamer-188 can attenuate blood-brain barrier damage to exert neuroprotective effect in mice intracerebral hemorrhage model. *J. Mol. Neurosci.*, 55(1):240–250, Jan 2015.

[100] K. Haman. *Development of Model Diblock Copolymer Surfactants for Mechanistic Investigations of Cell Membrane Stabilization.* PhD thesis, University of Minnesota, 2015.

[101] E. M. Houang, K. J. Haman, M. Kim, W. Zhang, D. A. Lowe, Y. Y. Sham, T. P. Lodge, B. J. Hackel, F. S. Bates, and J. M. Metzger. Chemical End

Group Modified Diblock Copolymers Elucidate Anchor and Chain Mechanism of Membrane Stabilization. *Mol. Pharm.*, 14(7):2333–2339, Jul 2017.

[102] A. Nystrom, K. Thriene, V. Mittapalli, J. S. Kern, D. Kiritsi, J. Dengjel, and L. Bruckner-Tuderman. Losartan ameliorates dystrophic epidermolysis bullosa and uncovers new disease mechanisms. *EMBO Mol Med*, 7(9):1211–1228, Sep 2015.

[103] E. Aikawa, R. Fujita, Y. Kikuchi, Y. Kaneda, and K. Tamai. Systemic high-mobility group box 1 administration suppresses skin inflammation by inducing an accumulation of PDGFRa+ mesenchymal cells from bone marrow. *Sci Rep*, 5:11008, Jun 2015.

[104] E. Marshall. Gene therapy death prompts review of adenovirus vector. *Science*, 286(5448):2244–2245, Dec 1999.

[105] J. Couzin and J. Kaiser. Gene therapy. As Gelsinger case ends, gene therapy suffers another blow. *Science*, 307(5712):1028, Feb 2005.

[106] S. Hacein-Bey-Abina, A. Garrigue, G. P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B. H. Belohradsky, U. Wintergerst, M. C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F. D. Bushman, A. Fischer, and M. Cavazzana-Calvo. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.*, 118(9):3132–3142, Sep 2008.

[107] H. C. J. Ertl and K. A. High. Impact of AAV Capsid-Specific T-Cell Responses on Design and Outcome of Clinical Gene Transfer Trials with Recombinant Adeno-Associated Viral Vectors: An Evolving Controversy. *Hum. Gene Ther.*, 28(4):328–337, 04 2017.

[108] F. Eichler, C. Duncan, P. L. Musolino, P. J. Orchard, S. De Oliveira, A. J. Thrasher, M. Armant, C. Dansereau, T. C. Lund, W. P. Miller, G. V. Raymond, R. Sankar, A. J. Shah, C. Sevin, H. B. Gaspar, P. Gissen, H. Amartino, D. Bratkovic, N. J. C. Smith, A. M. Paker, E. Shamir, T. O'Meara, D. Davidson, P. Aubourg, and D. A. Williams. Hematopoietic Stem-Cell Gene Therapy for Cerebral Adrenoleukodystrophy. *N. Engl. J. Med.*, 377(17):1630–1638, 10 2017.

[109] J. A. Ribeil, S. Hacein-Bey-Abina, E. Payen, A. Magnani, M. Semeraro, E. Magrin, L. Caccavelli, B. Neven, P. Bourget, W. El Nemer, P. Bartolucci, L. Weber, H. Puy, J. F. Meritet, D. Grevent, Y. Beuzard, S. Chretien, T. Lefebvre, R. W. Ross, O. Negre, G. Veres, L. Sandler, S. Soni, M. de Montalembert, S. Blanche, P. Leboulch, and M. Cavazzana. Gene Therapy in a Patient with Sickle Cell Disease. *N. Engl. J. Med.*, 376(9):848–855, 03 2017.

[110] S. Rangarajan, L. Walsh, W. Lester, D. Perry, B. Madan, M. Laffan, H. Yu, C. Vettermann, G. F. Pierce, W. Y. Wong, and K. J. Pasi. AAV5-Factor VIII Gene Transfer in Severe Hemophilia A. *N. Engl. J. Med.*, 377(26):2519–2530, 12 2017.

[111] L. A. George, S. K. Sullivan, A. Giermasz, J. E. J. Rasko, B. J. Samelson-Jones, J. Ducore, A. Cuker, L. M. Sullivan, S. Majumdar, J. Teitel, C. E. McGuinn, M. V. Ragni, A. Y. Luk, D. Hui, J. F. Wright, Y. Chen, Y. Liu, K. Wachtel, A. Winters, S. Tiefenbacher, V. R. Arruda, J. C. M. van der Loo, O. Zelenaia, D. Takefman, M. E. Carr, L. B. Couto, X. M. Anguela, and K. A. High. Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. *N. Engl. J. Med.*, 377(23):2215–2227, 12 2017.

[112] S. Russell, J. Bennett, J. A. Wellman, D. C. Chung, Z. F. Yu, A. Tillman, J. Wittes, J. Pappas, O. Elci, S. McCague, D. Cross, K. A. Marshall, J. Walshire, T. L. Kehoe, H. Reichert, M. Davis, L. Raffini, L. A. George, F. P. Hudson, L. Dingfield, X. Zhu, J. A. Haller, E. H. Sohn, V. B. Mahajan,

W. Pfeifer, M. Weckmann, C. Johnson, D. Gewaily, A. Drack, E. Stone, K. Wachtel, F. Simonelli, B. P. Leroy, J. F. Wright, K. A. High, and A. M. Maguire. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet*, 390(10097):849–860, Aug 2017.

[113] J. Kaiser. A second chance. *Science*, 358(6363):582–585, Nov 2017.

[114] J. R. Mendell, S. Al-Zaidy, R. Shell, W. D. Arnold, L. R. Rodino-Klapac, T. W. Prior, L. Lowes, L. Alfano, K. Berry, K. Church, J. T. Kissel, S. Nagendran, J. L'Italien, D. M. Sproule, C. Wells, J. A. Cardenas, M. D. Heitzer, A. Kaspar, S. Corcoran, L. Braun, S. Likhite, C. Miranda, K. Meyer, K. D. Foust, A. H. M. Burghes, and B. K. Kaspar. Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *N. Engl. J. Med.*, 377(18):1713–1722, 11 2017.

[115] M. J. Osborn, C. G. Starker, A. N. McElroy, B. R. Webber, M. J. Riddle, L. Xia, A. P. DeFeo, R. Gabriel, M. Schmidt, C. von Kalle, D. F. Carlson, M. L. Maeder, J. K. Joung, J. E. Wagner, D. F. Voytas, B. R. Blazar, and J. Tolar. TALEN-based gene correction for epidermolysis bullosa. *Mol. Ther.*, 21(6):1151–1159, Jun 2013.

[116] W. E. Lowry, L. Richter, R. Yachechko, A. D. Pyle, J. Tchieu, R. Sridharan, A. T. Clark, and K. Plath. Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl. Acad. Sci. U.S.A.*, 105(8):2883–2888, Feb 2008.

[117] J. Tolar, L. Xia, M. J. Riddle, C. J. Lees, C. R. Eide, R. T. McElmurry, M. Titeux, M. J. Osborn, T. C. Lund, A. Hovnanian, J. E. Wagner, and B. R. Blazar. Induced pluripotent stem cells from individuals with recessive dystrophic epidermolysis bullosa. *J. Invest. Dermatol.*, 131(4):848–856, Apr 2011.

[118] B. Munsky, G. Neuert, and A. van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, Apr 2012.

[119] K. E. Neu, Q. Tang, P. C. Wilson, and A. A. Khan. Single-Cell Genomics: Approaches and Utility in Immunology. *Trends Immunol.*, 38(2):140–149, 02 2017.

[120] D. Hebenstreit. Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology (Basel)*, 1(3):658–667, Nov 2012.

[121] R. Bacher and C. Kendziorski. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, 17:63, Apr 2016.

[122] De Wang, Feiping Nie, and Heng Huang. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*, ECMLPKDD'14, pages 306–321, Berlin, Heidelberg, 2014. Springer-Verlag.

[123] S. Islam, U. Kjallquist, A. Moliner, P. Zajac, J. B. Fan, P. Lonnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167, Jul 2011.

[124] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, Jun 2014.

[125] A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park,

and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, Jun 2013.

[126] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, 20(9):1131–1139, Sep 2013.

[127] A. Zeisel, A. B. Munoz-Manchado, S. Codeluppi, P. Lonnerberg, G. La Manno, A. Jureus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, Mar 2015.

[128] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015.

[129] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

[130] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, 32(4):381–386, Apr 2014.

[131] T. Kouno, M. de Hoon, J. C. Mar, Y. Tomaru, M. Kawano, P. Carninci, H. Suzuki, Y. Hayashizaki, and J. W. Shin. Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome Biol.*, 14(10):R118, 2013.

[132] C. Xu and Z. Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, Jun 2015.

[133] E. Pierson and C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, 16:241, Nov 2015.

[134] D. A. duVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics*, 17(1):363, Sep 2016.

[135] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5):495–502, May 2015.

[136] V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, and M. Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, 14(5):483–486, May 2017.

[137] S. C. Hicks, M. Teng, and Rafael A Irizarry. On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *bioRxiv*, 2015, https://www.biorxiv.org/content/early/2015/08/25/025528.full.pdf.

[138] A. A. Kolodziejczyk, J. K. Kim, J. C. Tsang, T. Ilicic, J. Henriksson, K. N. Natarajan, A. C. Tuck, X. Gao, M. Buhler, P. Liu, J. C. Marioni, and

S. A. Teichmann. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, 17(4):471–485, Oct 2015.

[139] B. Treutlein, D. G. Brownfield, A. R. Wu, N. F. Neff, G. L. Mantalas, F. H. Espinoza, T. J. Desai, M. A. Krasnow, and S. R. Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, May 2014.

[140] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, 8:14049, Jan 2017.

[141] J. A. Blake, J. T. Eppig, J. A. Kadin, J. E. Richardson, C. L. Smith, and C. J. Bult. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.*, 45(D1):D723–D729, Jan 2017.

[142] S. K. Mathai, B. S. Pedersen, K. Smith, P. Russell, M. I. Schwarz, K. K. Brown, M. P. Steele, J. E. Loyd, J. D. Crapo, E. K. Silverman, D. Nickerson, T. E. Fingerlin, I. V. Yang, and D. A. Schwartz. Desmoplakin Variants Are Associated with Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.*, 193(10):1151–1160, May 2016.

[143] P. N. Tsao, C. Matsuoka, S. C. Wei, A. Sato, S. Sato, K. Hasegawa, H. K. Chen, T. Y. Ling, M. Mori, W. V. Cardoso, and M. Morimoto. Epithelial Notch signaling regulates lung alveolar morphogenesis and airway epithelial integrity. *Proc. Natl. Acad. Sci. U.S.A.*, 113(29):8242–8247, 07 2016.

[144] S. Iinuma, E. Aikawa, K. Tamai, R. Fujita, Y. Kikuchi, T. Chino, J. Kikuta, J. A. McGrath, J. Uitto, M. Ishii, H. Iizuka, and Y. Kaneda. Transplanted bone marrow-derived circulating PDGFRa+ cells restore type VII collagen in recessive dystrophic epidermolysis bullosa mouse skin graft. *J. Immunol.*, 194(4):1996–2003, Feb 2015.

[145] G. Petrof, A. Abdul-Wahab, L. Proudfoot, R. Pramanik, J. E. Mellerio, and J. A. McGrath. Serum levels of high mobility group box 1 correlate with disease severity in recessive dystrophic epidermolysis bullosa. *Exp. Dermatol.*, 22(6):433–435, Jun 2013.

[146] K. Tamai, T. Yamazaki, T. Chino, M. Ishii, S. Otsuru, Y. Kikuchi, S. Iinuma, K. Saga, K. Nimura, T. Shimbo, N. Umegaki, I. Katayama, J. Miyazaki, J. Takeda, J. A. McGrath, J. Uitto, and Y. Kaneda. PDGFRa+ cells in bone marrow are mobilized by high mobility group box 1 (HMGB1) to regenerate injured epithelia. *Proc. Natl. Acad. Sci. U.S.A.*, 108(16):6609–6614, Apr 2011.

[147] D. Anastassiou. Comment on "A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets," Cancer Lett. 2016 Nov 28; 382 (2): 203-214. *Cancer Lett.*, 393:125–126, 05 2017.

[148] D. Staloch, X. Gao, K. Liu, M. Xu, X. Feng, J. F. Aronson, M. Falzon, G. H. Greeley, C. Rastellini, C. Chao, M. R. Hellmich, Y. Cao, and T. C. Ko. Gremlin is a key pro-fibrogenic factor in chronic pancreatitis. *J. Mol. Med.*, 93(10):1085–1093, Oct 2015.

[149] R. H. Church, I. Ali, M. Tate, D. Lavin, A. Krishnakumar, H. M. Kok, J. R. Hombrebueno, P. D. Dunne, V. Bingham, R. Goldschmeding, F. Martin, and D. P. Brazil. Gremlin1 plays a key role in kidney development and renal fibrosis. *Am. J. Physiol. Renal Physiol.*, 312(6):F1141–F1157, Jun 2017.

[150] M. D. Combs, R. H. Knutsen, T. J. Broekelmann, H. M. Toennies, T. J. Brett, C. A. Miller, D. L. Kober, C. S. Craft, J. J. Atkinson, J. M. Shipley, B. C. Trask, and R. P. Mecham. Microfibril-associated glycoprotein 2 (MAGP2) loss of function has pleiotropic effects in vivo. *J. Biol. Chem.*, 288(40):28869–28880, Oct 2013.

[151] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[152] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans Neural Netw*, 22(2):199–210, Feb 2011.

[153] Zheng Wang, Yangqiu Song, and Changshui Zhang. Transferred dimensionality reduction. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 550–565, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[154] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 200–207, New York, NY, USA, 2008. ACM.

[155] O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, 16(3):133–145, Mar 2015.

[156] K. K. Dey, C. J. Hsiao, and M. Stephens. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.*, 13(3):e1006599, Mar 2017.

[157] B. A. Lindborg, J. H. Brekke, A. L. Vegoe, C. B. Ulrich, K. T. Haider, S. Subramaniam, S. L. Venhuizen, C. R. Eide, P. J. Orchard, W. Chen, Q. Wang, F. Pelaez, C. M. Scott, E. Kokkoli, S. A. Keirstead, J. R. Dutton, J. Tolar, and T. D. O'Brien. Rapid Induction of Cerebral Organoids From

Human Induced Pluripotent Stem Cells Using a Chemically Defined Hydrogel and Defined Cell Culture Medium. *Stem Cells Transl Med*, 5(7):970–979, Jul 2016.

[158] A. Kramer, J. Green, J. Pollard, and S. Tugendreich. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530, Feb 2014.

[159] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[160] M. A. Hillmyer and F. S. Bates. Synthesis and Characterization of Model Polyalkane-Poly(ethylene oxide) Block Copolymers. *Macromolecules*, 29(22):6994–7002, 1996.

[161] S. Ndoni. Laboratory-scale setup for anionic polymerization under inert atmosphere. *Review of Scientific Instruments*, 66(2):1090, feb 1995.

[162] Jifeng Ding, Colin Price, and Colin Booth. Use of crown ether in the anionic polymerization of propylene oxide-1. rate of polymerization. *European Polymer Journal*, 27(9):891 – 894, 1991.

[163] Jifeng Ding, Frank Heatley, Colin Price, and Colin Booth. Use of crown ether in the anionic polymerization of propylene oxide-2. molecular weight and molecular weight distribution. *European Polymer Journal*, 27(9):895 – 899, 1991.

[164] Jifeng Ding, David Attwood, Colin Price, and Colin Booth. Use of crown ether in the anionic polymerization of propylene oxide-3. preparation and

micellization of diblock-copoly(oxypropylene/oxyethylene). *European Polymer Journal*, 27(9):901 – 905, 1991.

[165] E. K. Hollmann, A. K. Bailey, A. V. Potharazu, M. D. Neely, A. B. Bowman, and E. S. Lippmann. Accelerated differentiation of human induced pluripotent stem cells to blood-brain barrier endothelial cells. *Fluids Barriers CNS*, 14(1):9, Apr 2017.

[166] J. L. Mantle, L. Min, and K. H. Lee. Minimum Transendothelial Electrical Resistance Thresholds for the Study of Small and Large Molecule Drug Transport in a Human in Vitro Blood-Brain Barrier Model. *Mol. Pharm.*, 13(12):4191–4198, 12 2016.

[167] M. A. Deli, C. S. Abraham, Y. Kataoka, and M. Niwa. Permeability studies on in vitro blood-brain barrier models: physiology, pathology, and pharmacology. *Cell. Mol. Neurobiol.*, 25(1):59–127, Feb 2005.

[168] P. J. Gaillard and A. G. de Boer. Relationship between permeability status of the blood-brain barrier and in vitro permeability coefficient of a drug. *Eur J Pharm Sci*, 12(2):95–102, Dec 2000.

[169] J. L. Madara. Regulation of the movement of solutes across tight junctions. *Annu. Rev. Physiol.*, 60:143–159, 1998.

[170] No authors listed. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, 45(D1):D331–D338, Jan 2017.

[171] K. R. Kam, L. A. Walsh, S. M. Bock, M. Koval, K. E. Fischer, R. F. Ross, and T. A. Desai. Nanostructure-mediated transport of biologics across epithelial tissue: enhancing permeability via nanotopography. *Nano Lett.*, 13(1):164–171, Jan 2013.

[172] H. Powell, R. Tindall, P. Schultz, D. Paa, J. O'Brien, and P. Lampert. Adrenoleukodystrophy. Electron microscopic findings. *Arch. Neurol.*, 32(4):250–260, Apr 1975.

[173] J. M. Powers and H. H. Schaumburg. Adreno-leukodystrophy (sex-linked Schilder's disease). A pathogenetic hypothesis based on ultrastructural lesions in adrenal cortex, peripheral nerve and testis. *Am. J. Pathol.*, 76(3):481–491, Sep 1974.

[174] H. H. Schaumburg, E. P. Richardson, P. C. Johnson, R. B. Cohen, J. M. Powers, and C. S. Raine. Schilder's disease. Sex-linked recessive transmission with specific adrenal changes. *Arch. Neurol.*, 27(5):458–460, Nov 1972.

[175] A. Schluter, L. Espinosa, S. Fourcade, J. Galino, E. Lopez, E. Ilieva, L. Morato, M. Asheuer, T. Cook, A. McLaren, J. Reid, F. Kelly, S. Bates, P. Aubourg, E. Galea, and A. Pujol. Functional genomic analysis unravels a metabolic-inflammatory interplay in adrenoleukodystrophy. *Hum. Mol. Genet.*, 21(5):1062–1077, Mar 2012.

[176] M. C. van de Beek, R. Ofman, I. Dijkstra, F. Wijburg, M. Engelen, R. Wanders, and S. Kemp. Lipid-induced endoplasmic reticulum stress in X-linked adrenoleukodystrophy. *Biochim. Biophys. Acta*, 1863(9):2255–2265, Sep 2017.

[177] C. Y. Cheng, J. Y. Wang, R. Kausik, K. Y. Lee, and S. Han. Nature of interactions between PEO-PPO-PEO triblock copolymers and lipid membranes: (II) role of hydration dynamics revealed by dynamic nuclear polarization. *Biomacromolecules*, 13(9):2624–2633, Sep 2012.

[178] J. Y. Wang, J. Marks, and K. Y. Lee. Nature of interactions between PEO-PPO-PEO triblock copolymers and lipid membranes: (I) effect of polymer hydrophobicity on its ability to protect liposomes from peroxidation. *Biomacromolecules*, 13(9):2616–2623, Sep 2012.

[179] C. Theda, K. Gibbons, T. E. Defor, P. K. Donohue, W. C. Golden, A. D. Kline, F. Gulamali-Majid, S. R. Panny, W. C. Hubbard, R. O. Jones, A. K. Liu, A. B. Moser, and G. V. Raymond. Newborn screening for X-linked

adrenoleukodystrophy: further evidence high throughput screening is feasible. *Mol. Genet. Metab.*, 111(1):55–57, Jan 2014.

[180] A. R. Kemper, J. Brosco, A. M. Comeau, N. S. Green, S. D. Grosse, E. Jones, J. M. Kwon, W. K. Lam, J. Ojodu, L. A. Prosser, and S. Tanksley. Newborn screening for X-linked adrenoleukodystrophy: evidence summary and advisory committee recommendation. *Genet. Med.*, 19(1):121–126, 01 2017.

[181] F. D. Weber, C. Wiesinger, S. Forss-Petter, G. Regelsberger, A. Einwich, W. H. Weber, W. Kohler, H. Stockinger, and J. Berger. X-linked adrenoleukodystrophy: very long-chain fatty acid metabolism is severely impaired in monocytes but not in lymphocytes. *Hum. Mol. Genet.*, 23(10):2542–2550, May 2014.

[182] N. Cartier, S. Hacein-Bey-Abina, C. C. Bartholomae, G. Veres, M. Schmidt, I. Kutschera, M. Vidaud, U. Abel, L. Dal-Cortivo, L. Caccavelli, N. Mahlaoui, V. Kiermer, D. Mittelstaedt, C. Bellesme, N. Lahlou, F. Lefrere, S. Blanche, M. Audit, E. Payen, P. Leboulch, B. l'Homme, P. Bougneres, C. Von Kalle, A. Fischer, M. Cavazzana-Calvo, and P. Aubourg. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science*, 326(5954):818–823, Nov 2009.

[183] P. J. Orchard and J. Tolar. Transplant outcomes in leukodystrophies. *Semin. Hematol.*, 47(1):70–78, Jan 2010.

[184] K. Miki, K. Endo, S. Takahashi, S. Funakoshi, I. Takei, S. Katayama, T. Toyoda, M. Kotaka, T. Takaki, M. Umeda, C. Okubo, M. Nishikawa, A. Oishi, M. Narita, I. Miyashita, K. Asano, K. Hayashi, K. Osafune, S. Yamanaka, H. Saito, and Y. Yoshida. Efficient Detection and Purification of Cell Populations Using Synthetic MicroRNA Switches. *Cell Stem Cell*, 16(6):699–711, Jun 2015.

[185] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10, 2011, ISSN 2226-6089.

[186] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Mar 2012.

[187] Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, Apr 2014.

[188] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42(Database issue):68–73, Jan 2014.

[189] M. Eisenstein. Oxford Nanopore announcement sets sequencing sector abuzz. *Nat. Biotechnol.*, 30(4):295–296, Apr 2012.

[190] E. Karlsson, A. Larkeryd, A. Sjodin, M. Forsman, and P. Stenberg. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci Rep*, 5:11996, Jul 2015.

[191] N. J. Loman, J. Quick, and J. T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*, 12(8):733–735, Aug 2015.

[192] S. A. Moschos. Ebola Check: Delivering molecular diagnostics at the point of need. *Hell J Nucl Med*, 18 Suppl 1:144, 2015.

[193] E. Check Hayden. Pint-sized DNA sequencer impresses first users. *Nature*, 521(7550):15–16, May 2015.

[194] R. Zilionis, J. Nainys, A. Veres, V. Savova, D. Zemmour, A. M. Klein, and L. Mazutis. Single-cell barcoding and sequencing using droplet microfluidics. *Nat Protoc*, 12(1):44–73, Jan 2017.

Appendix A

Supplementary Information

A.1 Supplementary algorithms

A.1.1 Supplementary chapter 2 algorithm

Minimizing Equation 2.3 For simplicity, the objective function in equation (2.3) is replaced by

$$\mathcal{J}(U^{(d)}) = \frac{1}{2}\mathcal{J}_1 - w\mathcal{J}_2 + \alpha\mathcal{J}_3. \tag{A.1}$$

Then the gradient of $U_{i,:}^{(d)T}$ is obtained as

$$\frac{\partial \mathcal{J}}{\partial (U_{i,:}^{(d)T})} = \frac{1}{2}\frac{\partial \mathcal{J}_1}{\partial (U_{i,:}^{(d)T})} - w\frac{\partial \mathcal{J}_2}{\partial (U_{i,:}^{(d)T})} + \alpha\frac{\partial \mathcal{J}_3}{\partial (U_{i,:}^{(d)T})}. \tag{A.2}$$

Let $\Psi = \mathbf{I}_k - \frac{\mathbf{1}_k \mathbf{1}_k^T}{k}$ and $\Phi = \mathbf{I}_d - \frac{\mathbf{1}_d \mathbf{1}_d^T}{d}$, then

$$\frac{\partial \mathcal{J}_1}{\partial (U_{i,:}^{(d)T})} = 2B_i V^{(d)T}(V^{(d)} U_{i,:}^{(d)T} - X_{i,:}^{(d)T}) \tag{A.3}$$

$$\frac{\partial \mathcal{J}_2}{\partial (U_{i,:}^{(d)T})} = \frac{2B_i}{k} \Psi U_{i,:}^{(d)T} \tag{A.4}$$

$$\frac{\partial \mathcal{J}_3}{\partial (U_{i,:}^{(d)T})} = \frac{2B_i k}{d}(\Phi_{d,d} U_{i,:}^{(d)T} + \sum_{l \neq d} \Phi_{dl} U_{i,:}^{(l)T}). \tag{A.5}$$

Finally, bringing equations (A.3), (A.4) and (A.5) back into (A.2) becomes

$$\begin{aligned} \frac{\partial \mathcal{J}}{\partial (U_{i,:}^{(d)T})} &= B_i(V^{(d)T} V^{(d)} - \frac{2w}{k}\Psi + \frac{2\alpha k \Phi_{d,d}}{d}\mathbf{I}_k) U_{i,:}^{(d)T} \\ &- B_i V^{(d)T} X_{i,:}^{(d)T} + \frac{2B_i \alpha k}{d} \sum_{l \neq d} \Phi_{dl} U_{i,:}^{(l)T}. \end{aligned} \tag{A.6}$$

When w and α are properly chosen (see section 2.2.3) and $B_i = 1$, objective (A.1) is convex. By setting the derivative (A.6) to zero the analytical solution is then

$$\begin{aligned} U_{i,:}^{(d)^T} = (V^{(d)^T} V^{(d)} - \frac{2w}{k}\Psi + \frac{2\alpha k \Phi_{d,d}}{d}\mathbf{I}_k)^{-1}(V^{(d)^T} X_{i,:}^{(d)^T} - \\ \frac{2\alpha k}{d} \sum_{l \neq d} \Phi_{dl} U_{i,:}^{(l)T}). \end{aligned} \tag{A.7}$$

A.2 Supplementary tables

A.2.1 Supplementary chapter 2 tables

S1 Table. RDEB patient and donor demographics. RDEB patient and HLA-matched sibling age and gender at the time of sample collection.

Pair	Patient Age	Patient Gender	Donor Age	Donor Gender
1	12	Female	9	Female
2	12	Female	13	Female
3	12	Male	12	Male
4	1	Female	3	Female
5	5	Male	1	Male
6	5	Female	14	Female

S2 Table. Primary antibodies for flow cytometry.

Target Antigen	Antibody species	Vendor	Product Number	Dilution
GREM-1	Rabbit	Cloud Clone	PAC128Hu02	1:200
COL11A1	Rabbit	Abcam	Ab64883	1:200
MFAP5/MAGP2	Rabbit	Abcam	Ab203828	1:200
Rabbit IgG	Rabbit	Jackson ImmunoResearch	011-000-003	1:200

S3 Table. Secondary antibodies used for flow cytometry.

Species Reactivity	Host	Conjugate	Vendor	Dilution
Rabbit	Goat	Alexa Fluor 594	ThermoFisher	1:200

A.2.2 Supplementary chapter 3 tables

iPSC lines	Sex	Derived cell type	Delivery method	Reprogramming factors
ccALD1	Male	Fibroblasts	Retrovirus	OCT4, SOX2, KLF4, c-MYC
ccALD2	Male	Fibroblasts	Retrovirus	OCT4, SOX2, KLF4, c-MYC
ccALD3	Male	Keratinocytes	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT1	Female	Keratinocytes	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT2	Male	Urine cells	Retrovirus	OCT4, SOX2, KLF4, c-MYC
WT3	Male	CD34+ bone marrow cells	Sendai virus	OCT4, SOX2, KLF4, c-MYC

S4 Table. Information on induced pluripotent stem cell (iPSC) lines used in study.

S5 Table. Primary antibodies used for immunocytochemistry.

Target antigen	Antibody species	Vendor	Clone or product number	Dilution
PECAM-1	Rabbit	ThermoFisher	RB10333P	1:100
GLUT-1	Mouse	ThermoFisher	MS10637P1; clone SPM498, IgG2a	1:100
Occludin	Mouse	ThermoFisher	331500; clone OC-3F10	1:200
Claudin-5	Mouse	ThermoFisher	352500; clone 4C3C2	1:50

S6 Table. Secondary antibodies used for immunocytochemistry.

Species reactivity	Host	Conjugate	Vendor	Dilution
Mouse	Goat	Alexa Fluor 594	ThermoFisher	1:200
Rabbit	Goat	Alexa Fluor 594	ThermoFisher	1:200

S7 Table. Secondary antibodies used for immunocytochemistry.

Gene Vendor	ID	Number
GAPDH	BioRad	qHsaCED0038674
CDH5	BioRad	qHsaCID0016288
SLC2A1	BioRad	qHsaCID0022232
ABCB1	BioRad	qHsaCID0020960

A.2.3 Supplementary chapter 4 tables

S8 Table. Differentially expressed (FDR ≤ 0.05) miRs between blistered and mosaic fibroblast populations following small RNA sequencing.

MicroRNA	Log_2(fold change)	P-val (FDR ≤ 0.05)
miR-#1	-2.21	1.4×10^{-16}
miR-#2	2.10	3.4×10^{-5}
miR-#3	-1.58	5.0×10^{-4}
miR-#4	-1.07	2.0×10^{-3}
miR-#5	-1.07	1.2×10^{-2}
miR-#6	1.36	2.0×10^{-2}
miR-#7	1.10	2.0×10^{-2}
miR-#8	0.95	2.0×10^{-2}
miR-#9	-2.15	2.1×10^{-2}
miR-#10	-0.88	2.5×10^{-2}

A.3 Supplementary figures

A.3.1 Supplementary chapter 2 figures

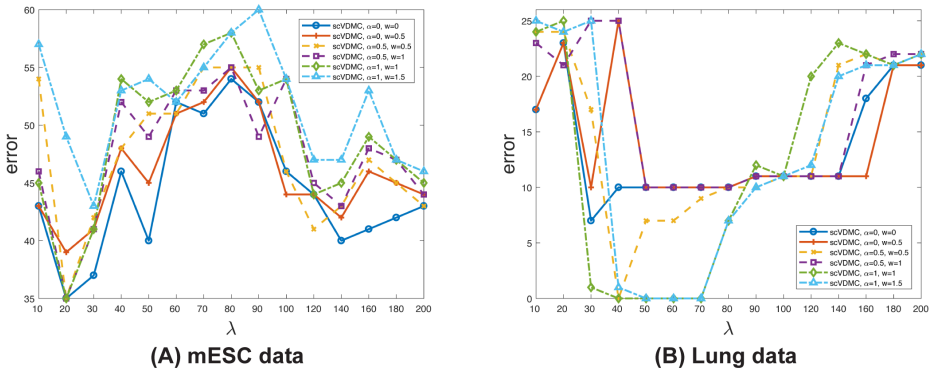


Figure A.1: **scVDMC clustering results under varying w on the mESC data and Lung data.**

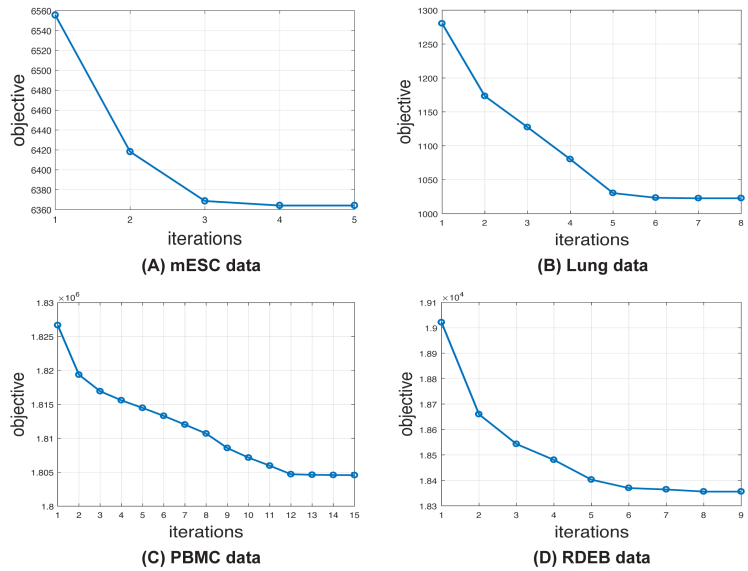


Figure A.2: **Convergence of scVDMC.** The object function in equation 2.1 is plotted under each iteration on the four datasets. In (A), (B) and (C), the parameters are $\alpha = 1, W = 0.1$ and $\lambda = 50$. In (D), the parameters are $\alpha = 1, W = 0.5$ and $\lambda = 300$, and the number of samples used is 1000 from donor A.

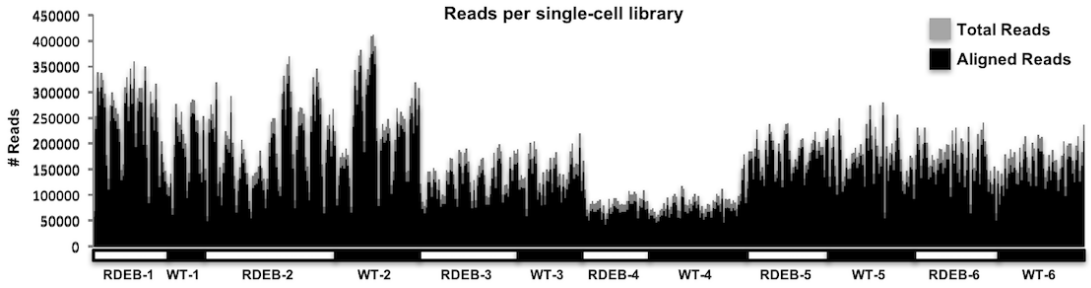


Figure A.3: **Read counts in the single cells.** The total number of the reads and the number of aligned reads are shown in each single-cell library. RDEB and WT individual pairs are indicated underneath.

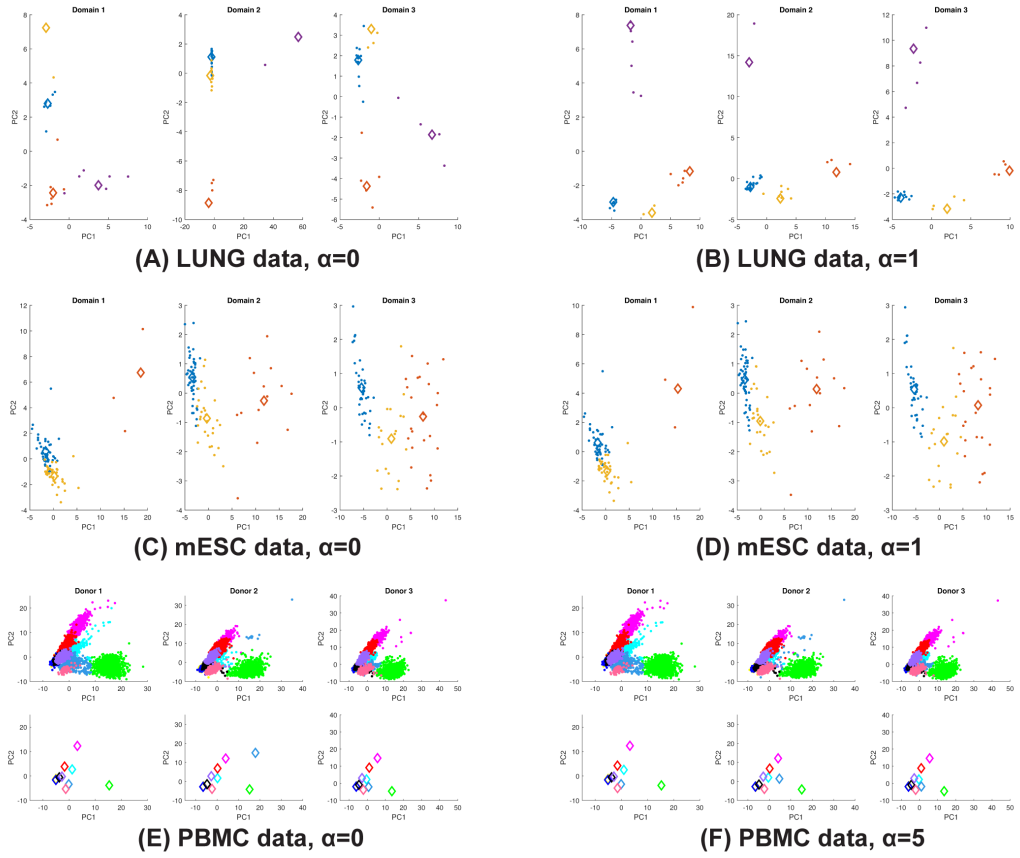


Figure A.4: **Capture of distinct single-cell populations by parameter tuning.** PCA is applied to the single cell profiles of the marker genes learned by scVDMC from the combined cell populations in each dataset. Each plot shows the projection of the data and the cluster centers by the first two principle components. The clusters are shown in different colors and the cluster centers are indicated by the diamonds . The projections with $\alpha = 0$ and 1 are compared on LUNG and mESC data and the projection with $\alpha = 0$ and 5 are compared on PBMC data. In (E) and (F), the data and the cluster centers are shown separately.

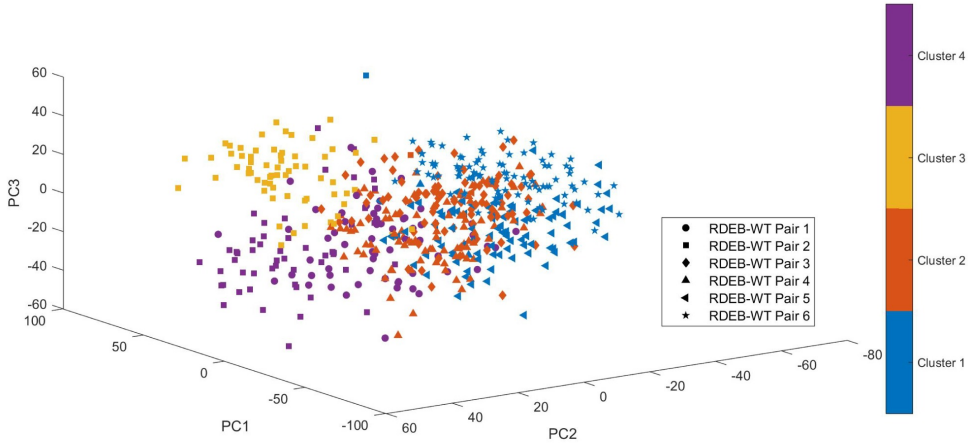


Figure A.5: **Pooled clustering of RDEB data with SC3.** SC3 was applied to cluster the single-cell populations from the six RDEB-WT pairs. PCA was applied to project the combined single cell profiles of the all genes from the pooled six cell populations in the first three PCs.

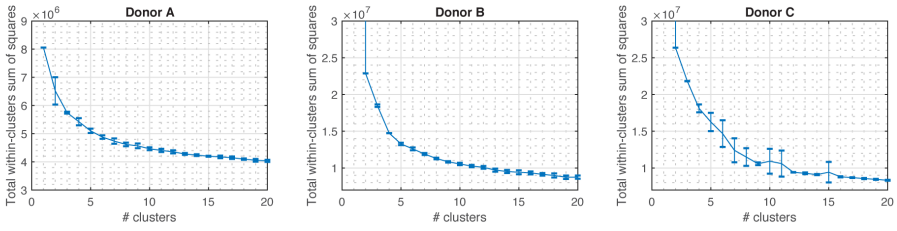


Figure A.6: **Determining the number of clusters in PBMC data with an "elbow" plot.** The mean total within-clusters sum of squares of the clustering averaged in ten repeats are shown for different choices of the number of clusters. The optimal number of clusters is around 10 in all the three donors.

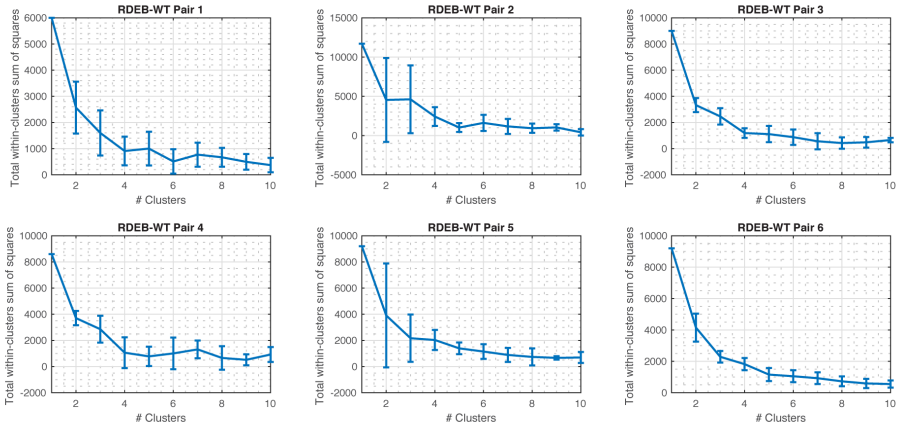


Figure A.7: **Determining the number of clusters in RDEB data with "elbow" plot.** The mean total within-clusters sum of squares of the clustering averaged in ten repeats are shown for different choices of the number of clusters. The "elbow" starts from 4 in all the six RDEB-WT pairs.

A.3.2 Supplementary chapter 3 figures

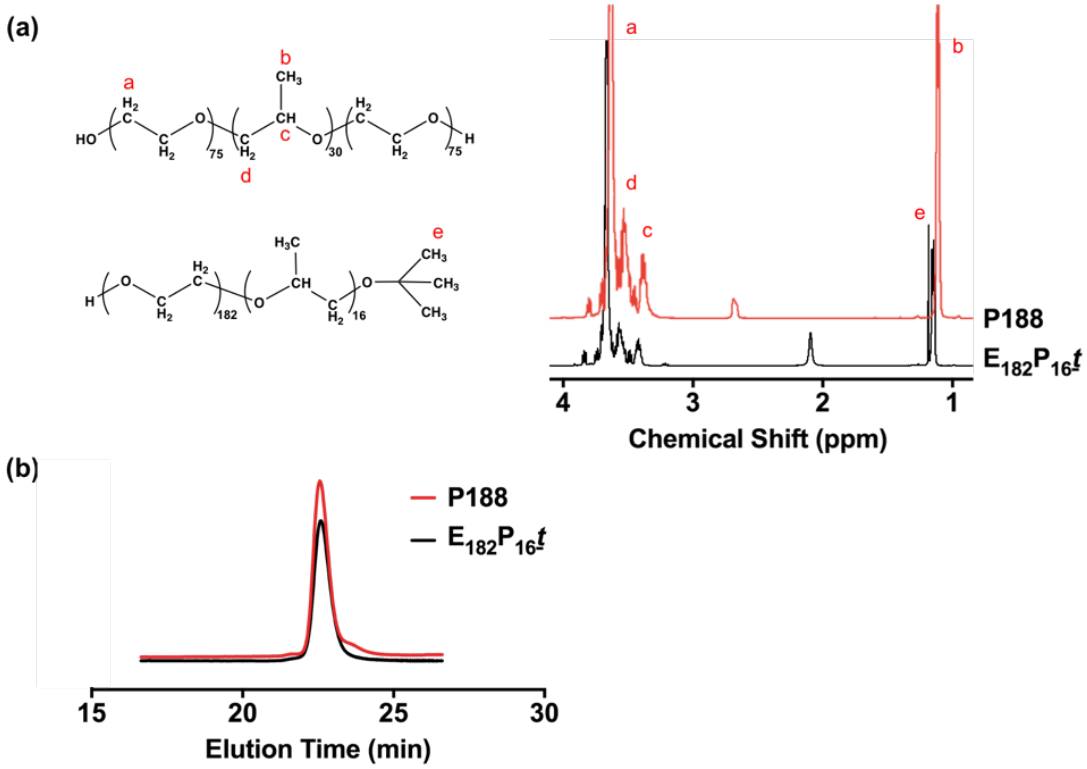


Figure A.8: **Polymer characterization data.** (a) $_1$H-NMR spectra of P188 and $E_{182}P_{16}\underline{t}$. (b) Size exclusion chromatograms of P188 and $E_{182}P_{16}\underline{t}$.

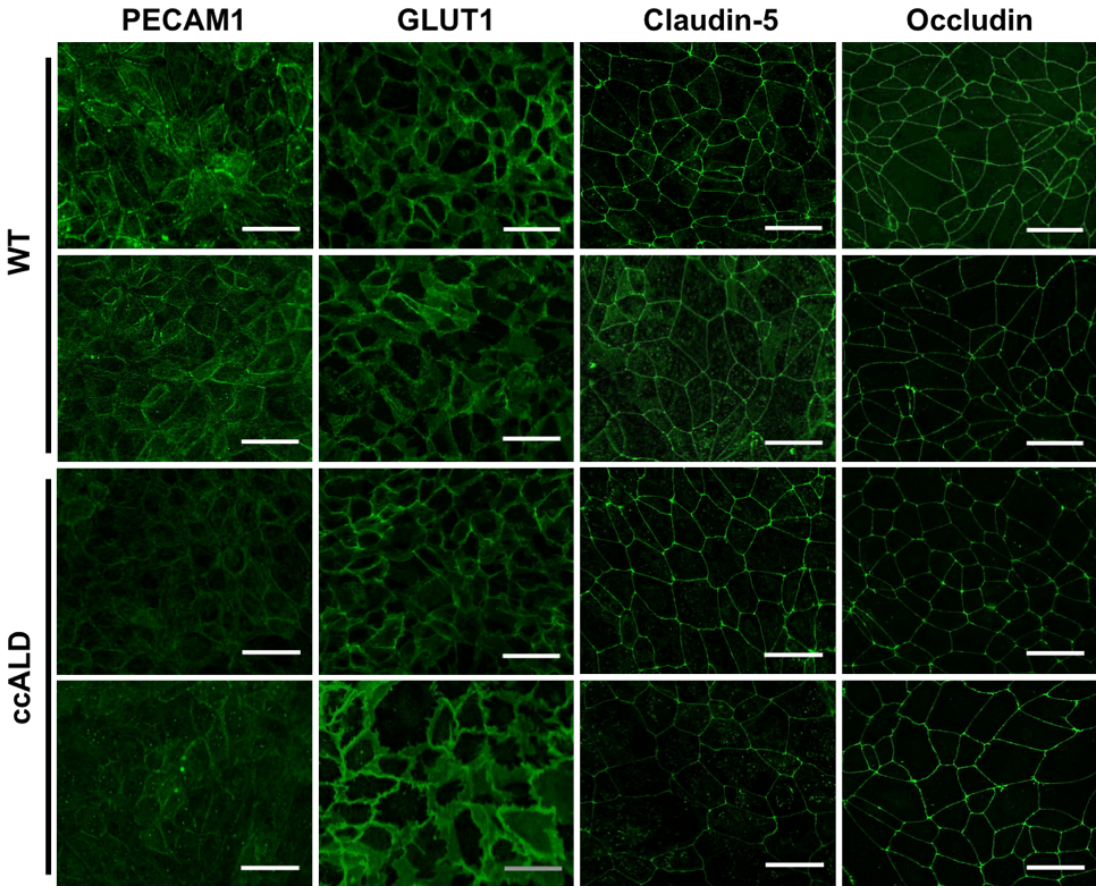


Figure A.9: **Representative immunocytochemistry images of iBMEC lines not shown in main manuscript.** All iBMECs express the requisite endothelial, tight junction, and BBB markers by immunocytochemistry. iBMECs from ccALD patients and WT controls express PECAM1, GLUT1, claudin-5, and occludin. Top to bottom: WT2, WT3, ccALD1, ccALD3. Scale bar = 50 μm.

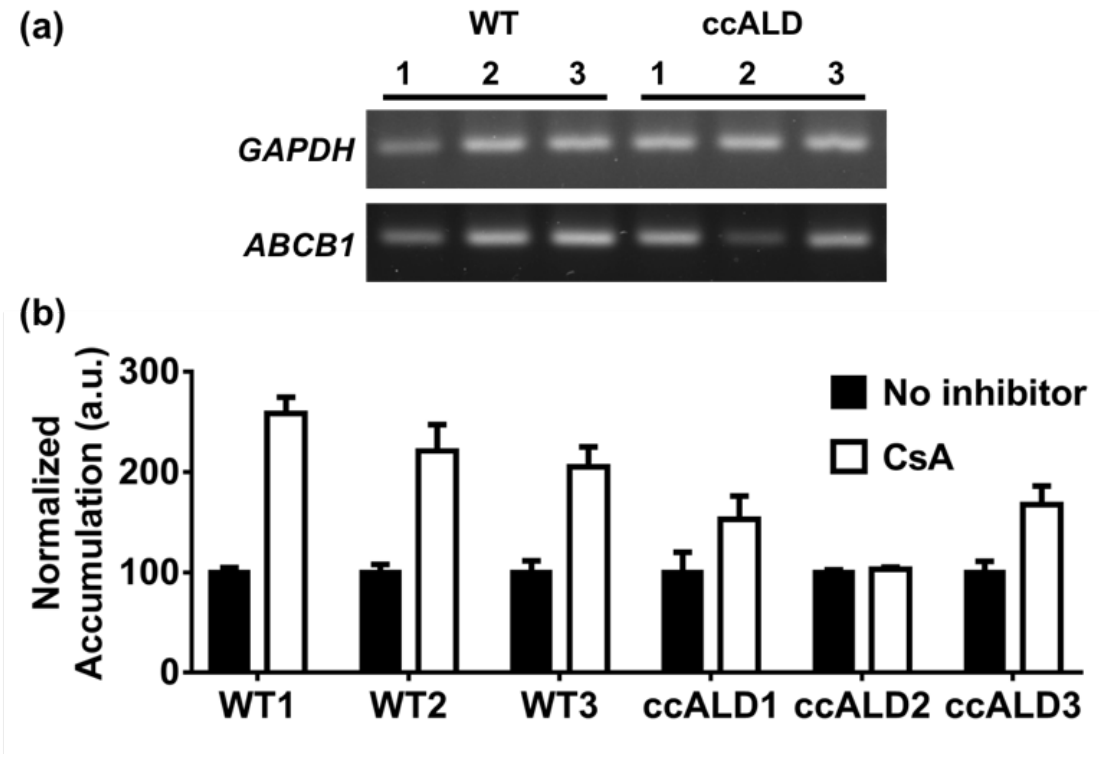


Figure A.10: **P-glycoprotein (P-gp) expression and function.** (a) All WT- and ccALD-iBMECs express *ABCB1*, which codes for BMEC-specific efflux transporter P-gp. (b) P-gp function was assessed with Rhodamine 123 accumulation assay, which showed functional P-gp for all iBMECs except ccALD2-iBMECs. Differences in normalized accumulation between no inhibitor and CsA inhibited samples are statistically significant ($p < 0.05$) for all cell lines with the exception of ccALD2-iBMECs. Fluorescence intensity is normalized by cell density, and accumulation is independently normalized to the corresponding control (no inhibitor). Four biological replicates used ($n = 4$).

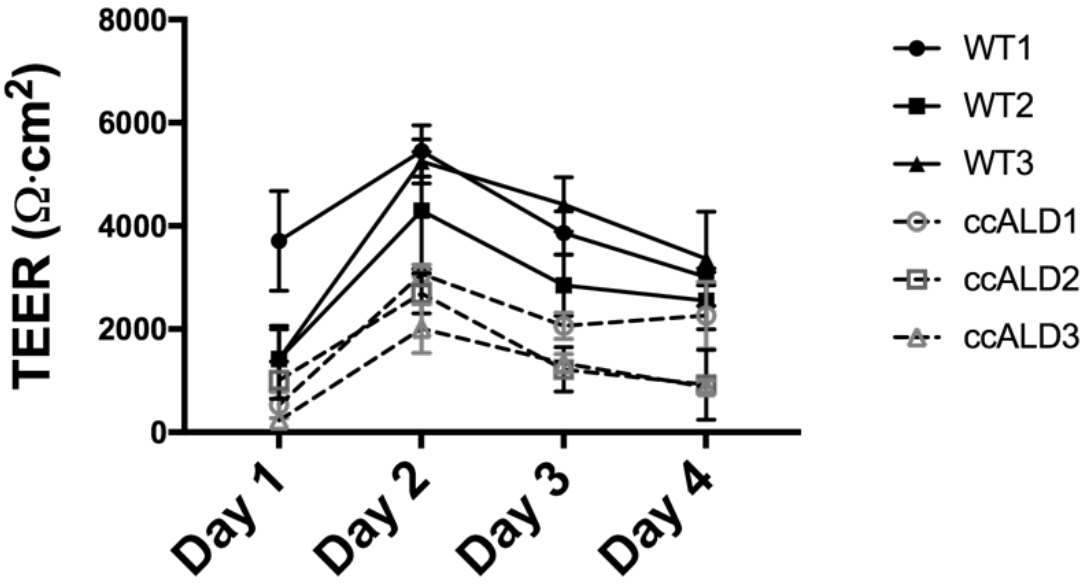


Figure A.11: **TEER measurements for individual cell lines.** Trans-endothelial electrical resistance (TEER) is lower for the ccALD-iBMECs compared to the WT-iBMECs at all days measured. Data compiled from three independent experiments with three biological replicates for each cell line (n = 9).

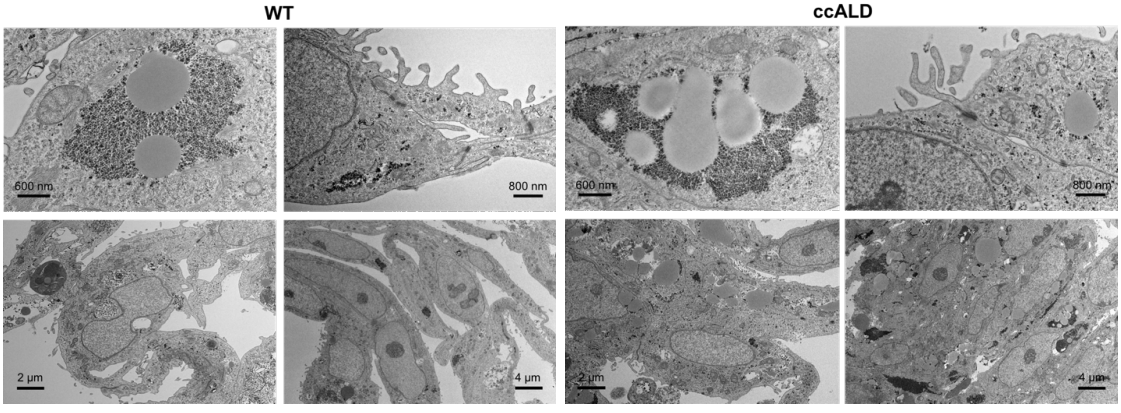


Figure A.12: **Additional representative TEM images.** TEM of WT1-iBMECs (above) and ccALD3-iBMECs (below) at varying magnifications showing increased lipid droplet accumulation in ccALD-iBMECs.

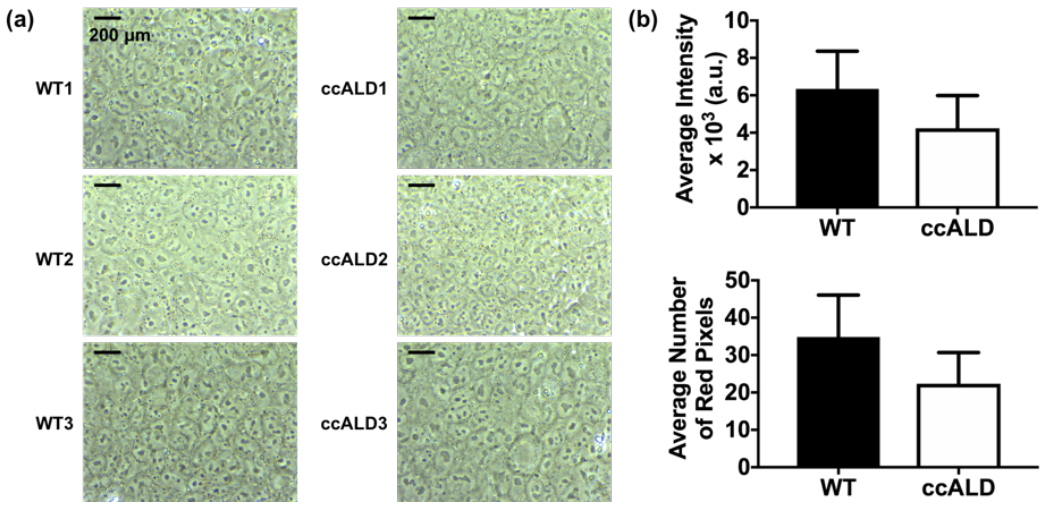


Figure A.13: **Oil-Red-O staining and quantification of WT and ccALD-iPSCs.** (a) Oil-Red-O staining of WT and ccALD-iPSCs show little to no lipid droplet accumulation. (b) Quantification of intensity and number of red pixels in images of Oil-Red-O stained iPSCs show no difference in lipid droplet accumulation in ccALD-iPSCs compared to WT-iPSCs. All cell lines were used for quantification with two biological replicates each (n=6).

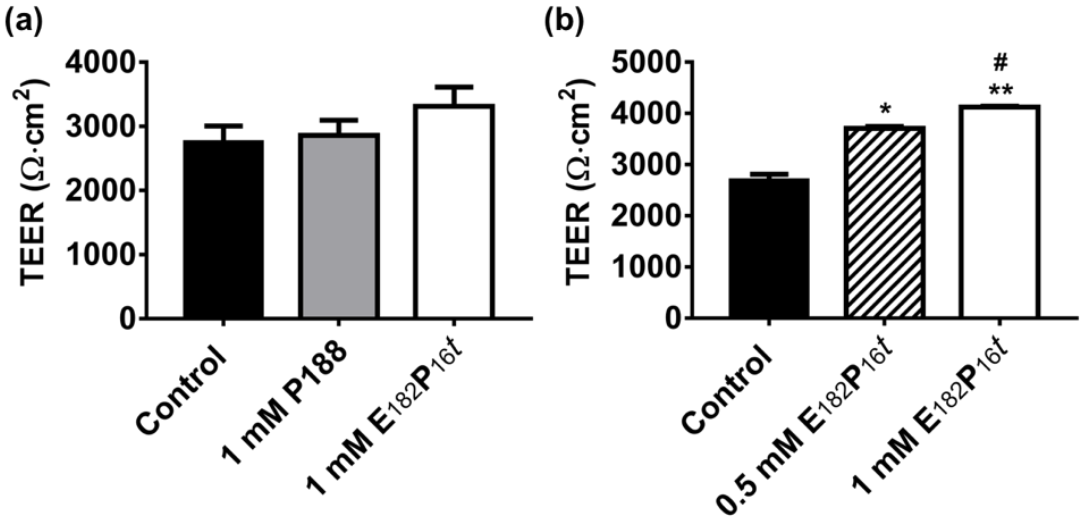


Figure A.14: **Timing and dosage effect of polymer treatment.** (a) Addition of 1 mM P188 or $E_{182}P_{16}\underline{t}$ on day 9 of the differentiation protocol had a non-significant effect on TEER of the ccALD3-iBMECs. However, a slight increase in TEER upon treatment with 1 mM $E_{182}P_{16}\underline{t}$ was observed. Data compiled from three independent experiments with three biological replicates each (n = 9). (b) Maximum TEER of ccALD3-iBMECs treated 1 mM $E_{182}P_{16}\underline{t}$ on day 3 of the differentiation protocol is higher than ccALD3-iBMECs treated with 0.5 mM $E_{182}P_{16}\underline{t}$ on day 3, signifying that treatment efficacy is concentration dependent. Data from three biological replicates (n = 3). * $p < 0.005$, ** $p < 0.0005$ with respect to control. # $p < 0.001$ with respect to 0.5 mM $E_{182}P_{16}\underline{t}$ condition.

A.3.3 Supplementary chapter 4 figures

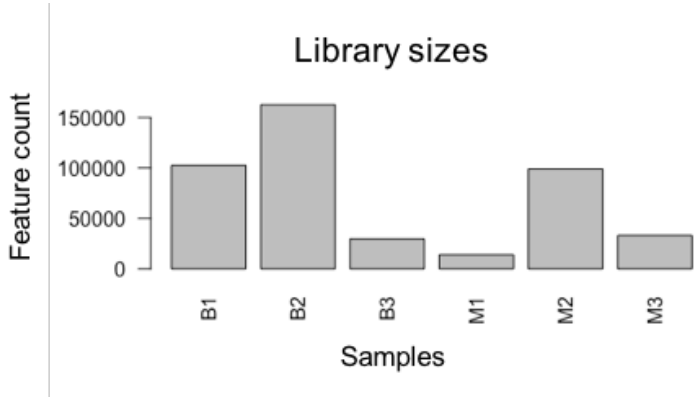


Figure A.15: **Library size measured in feature counts for each of the six samples used in the analysis.** Blistered: B1, B2, B3 and mosaic: M1, M2, M3 fibroblast populations.

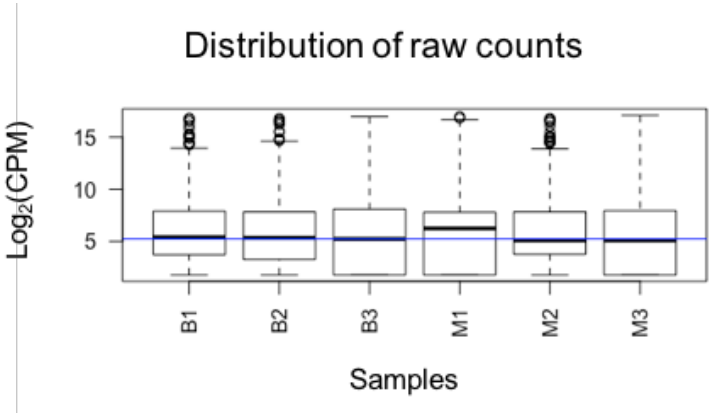


Figure A.16: **Distribution of raw counts after log transformation of counts per million (CPM) values for each of the six samples used in the analysis.** Blistered: B1, B2, B3 and mosaic: M1, M2, M3 fibroblast populations.

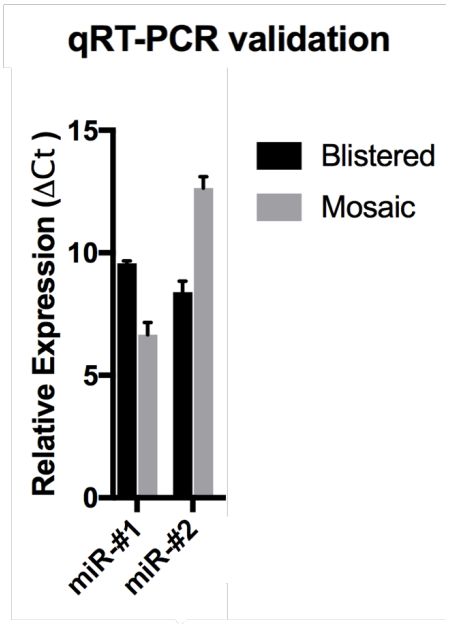


Figure A.17: **Quantitative RT-PCR analysis was performed to confirm miR expression patterns observed from small RNA-sequencing.** miR expression levels (ΔC_t) determined relative to U6 are represented. Mean ± SD shown. Data analyzed from three technical replicates (n = 3).

A.3.4 Supplementary chapter 5 figures

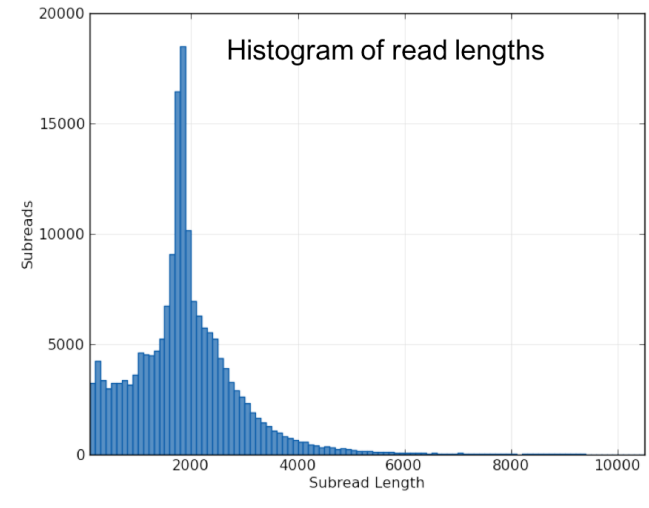


Figure A.18: **PacBio sequencing read lengths.** Histogram of read lengths.