Deep Learning for Skin Photoaging

Gokul Srinivasan
*Dartmouth College*, gokul.srinivasan.23@dartmouth.edu

Follow this and additional works at: https://digitalcommons.dartmouth.edu/cs_senior_theses

Part of the Computer Sciences Commons

**Recommended Citation**


This Thesis (Undergraduate) is brought to you for free and open access by the Computer Science at Dartmouth Digital Commons. It has been accepted for inclusion in Computer Science Senior Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.
DEEP LEARNING FOR SKIN PHOTOAGING

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Bachelors of Arts

in

Computer Science

by

Gokul Srinivasan

DARTMOUTH COLLEGE

Hanover, New Hampshire

May 2023
Abstract

Skin photoaging is the premature aging of skin that results from ultraviolet light exposure. It is a major risk factor for the development of skin cancer, among other malignant skin pathologies. Accordingly, understanding its etiology is important for both preventative and reparative clinical action. In this study, skin samples obtained from patients with ranging solar elastosis grades – a proxy for skin photoaging – were sequenced using next-generation sequencing techniques to further understand the genomic, epigenomic, and histological signs and signals of skin photoaging. The results of this study suggest that tissues with severe photoaging exhibit increases in the frequency of some immune cell populations, especially CD4, CD8, and regulatory T cells, macrophages, and mast cells, and decreases in the frequency of other immune cell populations, like dendritic cells, NKT cells, and plasma cells. Samples with severe solar elastosis also had increased expression of genes associated with the complement cascade and broader innate immune system. Methods to infer spatial transcriptomics data using purely histomorphological data were also devised in this work. Vision Transformers were trained to infer spatially variable gene expression across dichotomized and continuous regression tasks, reaching a median average AUC of 0.80 on the dichotomized task and a median average Spearman coefficient of 0.60 on the continuous task. Finally, methods were devised to spatially resolve chromatin accessibility data across whole slide images, and models were trained to predict spatial chromatin accessibility, reaching a median AUC score of 0.71. These results suggest that deep learning can be used to democratize access to spatial transcriptomics insights, facilitating a deeper understanding of disease etiologies informed by cell and tissue spatial dynamics.
I would like to express my gratitude to my thesis advisors, Professors Joshua Levy and Sarah Preum. Their sustained support and guidance were instrumental in completing this thesis. I would also like to thank Dr. Matt Davis for his continuous encouragement and clinical insight throughout this process.

A special note of thanks goes out to my friends. To Zach, Kamren, Alex, and Finn – know that this thesis would not have been possible without your support.

Finally, I want to extend my heartfelt thanks to my parents, Jaya and Shashi, and sister, Neha. My accomplishments – whatever their magnitude – belong to them.
## Contents

Abstract ................................................................. ii
Acknowledgements .................................................... iii

1 Introduction ......................................................... 1

2 Related Work ....................................................... 5

3 Methodology ........................................................ 10

3.1 Data Collection .................................................. 10

3.2 Dataset Description ............................................. 10

3.2.1 Image Data .................................................. 10

3.2.2 Genomics Data ............................................... 11

3.3 Tools and Libraries ............................................. 14

3.3.1 SeVI ....................................................... 14

3.3.2 Scanpy ...................................................... 14

3.3.3 Muon ......................................................... 16

3.3.4 Tangram ..................................................... 16

3.3.5 PyTorch ...................................................... 16

3.4 Computational Analyses ........................................ 16

3.4.1 Single Cell RNA Analysis .................................. 16

3.4.2 Single Cell ATAC Analysis ................................. 18

3.4.3 Visium Spatial Gene Expression Analysis ................. 19
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.4</td>
<td>Single Cell Multiome Analysis</td>
<td>19</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Virtual Spatial Gene Expression</td>
<td>20</td>
</tr>
<tr>
<td>4.1</td>
<td>Single Cell RNA Sequencing</td>
<td>25</td>
</tr>
<tr>
<td>4.2</td>
<td>Single Cell Multiome</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td>Virtual Spatial Gene Expression</td>
<td>31</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Spatial RNA Expression Prediction</td>
<td>31</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Spatial ATAC Prediction</td>
<td>43</td>
</tr>
<tr>
<td>5.0.1</td>
<td>ScRNA-seq Analysis</td>
<td>49</td>
</tr>
<tr>
<td>5.0.2</td>
<td>Multi-omics Analysis</td>
<td>51</td>
</tr>
<tr>
<td>5.0.3</td>
<td>Spatial RNA Prediction</td>
<td>52</td>
</tr>
<tr>
<td>5.0.4</td>
<td>Spatial Chromatin Accessibility Mapping and Prediction</td>
<td>56</td>
</tr>
</tbody>
</table>

5 Discussion

6 Supplementary Materials

References
Chapter 1

Introduction

Skin photoaging is the premature aging of skin attributable to ultraviolet radiation exposure (UVA and UVB) from the sun. While it remains difficult to precisely delineate the boundary between normal aging processes and skin photoaging, general symptoms of this affliction include loss of skin elasticity, lines and wrinkles of increasing number and depth, and pigment changes such as colored splotches and freckling [17]. In the later stages of severe skin photoaging, benign actinic keratoses, malignant squamous, and basal cell carcinoma tumors may develop [17].

While several different histomorphologic features have been studied as potential correlates for solar elastosis and other forms of skin aging, few of these correlations have been confirmed across multiple studies [17] [37]. Some reliable measures include epidermal thickness, the size of the dermal-epidermal junction, and increases in the presence of inflammatory markers [17]. Outside of histomorphology, genetic markers have demonstrated promise, a few of which – e.g., TP53, HES1, and KLF6 – have been more robustly associated with skin photoaging using both immunohistochemistry (IHC) and next-generation sequencing (NGS) [43][17]. In particular, NGS techniques such as single-cell RNA sequencing (scRNA-seq) and single-cell assay for transposase-accessible chromatin (scATAC-seq) work by fragmenting whole tissue samples into individual cells, and characterizing genetic and epigenetic signatures at the cell level. These techniques have greatly enhanced our understanding of the genetic underpinnings of skin aging. For example,
Sole-Boldo et al. use scRNA-seq to analyze the transcriptome of several thousand fibroblasts, concluding that partial loss of cellular identity (de-differentiation) is an important age-related change in the human dermis [32]. In a similar vein, Zou et al. leveraged scRNA-seq to characterize human skin transcriptomes [43] through age. They found alterations in key transcription factors – HES1 for fibroblasts and KLF6 for keratinocytes – that increased cell senescence and decreased cell proliferation[43]. These findings provide starting points for deeper research into the specific mechanisms and pathways these genes moderate, as well as targets for potential therapeutic intervention.

What single-cell sequencing techniques gain in specificity, they often lose in spatial connectedness. The vast majority of accessible scRNA-seq and scATAC-seq techniques work by decomposing tissues into individual cells, resulting in a loss of information on spatial relationships among the cells in a tissue sample. Fortunately, a range of techniques have been developed with the aim of spatial transcriptome mapping – that is, capturing gene expression at spatially localized points across a sample. These techniques allow researchers to uncover not only which, and to what extent, genes are expressed in a sample, but also where these genes are expressed spatially. This information can be especially useful when characterizing biological architectures that have important spatial dependencies, such as the tumor microenvironment or cell-cell interactions during ontogenesis [35]. For instance, Theocharidis et al. leveraged spatial transcriptomics to compare spatial gene expression among patients with and without healing diabetic foot ulcers. The healing group expressed a healing-associated fibroblast marker at sites adjacent to the wound bed, while the non-healing group lacked this gene localization [36].

While scRNA-seq and scATAC-seq techniques have been applied to the study of skin aging and photoaging, the area of research applying spatial transcriptomics to questions of skin photoaging is underexplored. Furthermore, few – if any – studies have understood skin photoaging using blended multi-omics assays, consisting of scRNA-seq, scATAC-seq, and spatial transcriptomics methods.
Together, these methods can yield single-cell resolution genetic and epigenetic information along-
side important spatial context.

As impressive as these multi-omics methods can be, they are often prohibitively expensive, costing
several thousands of dollars per sample. Spatial transcriptomics assays are particularly expensive,
and using these assays can be time and labor-intensive. To this end, recent efforts have been made
to infer spatial transcriptomics information using purely histomorphologic data [9]. For example,
Fatemi et al. leveraged deep learning models to accurately predict spatially variable gene expres-
sion in H&E-stained whole slide images (WSI) produced from colorectal biopsies[9]. However,
these models currently lack generalizability. A model trained to predict gene expression on col-
orectal WSI will not be able to accurately predict gene expression in skin samples. Therefore, there
is a need to develop deep-learning models unique to distinct organ systems.

Additionally, current spatial transcriptomics methods focus on spatial genome localization rather
than spatial epigenome localization. While recent research has attempted to extend the spatial
paradigm to measures of chromatin accessibility, these methods remain in their infancy and have
not yet been commercialized [7]. These methods, however, have yielded great insight, enabling
profiling of regulatory elements in the human tongue and developing mouse embryo [7]. Accord-
ingly, a simple and accessible method to acquire spatial chromatin accessibility information is
needed.

In this thesis, I attempt to further our understanding of the genetic and epigenetic basis of skin
photoaging, developing a range of novel genetic and epigenetic tools in the process. While a rich
body of literature exists regarding general skin aging, research into the etiology of skin photoaging
is more limited [32][43]. Even fewer studies utilize spatial transcriptomics to make progress on
the question of skin photoaging. In part, these trends result from barriers associated with NGS
 technique use more broadly, especially with respect to cost and skilled labor requirements. As
such, the main contributions of this thesis may be summarized as follows:

(a) Preliminary delineation of the genetic and histological changes correlated with skin photoaging using a multi-omics approach (i.e., a combination of scRNA-seq, scATAC-seq, and Visium spatial transcriptomics.)

(b) Development of novel spatial transcriptomics mapping and prediction tools for both gene expression and chromatin accessibility.

The remainder of this thesis will be structured as follows. Chapter 2 will situate my research in the context of related works. Chapter 3 will elaborate on the tools and methods used. Chapter 4 will delineate subsequent results. Finally, chapter 5 will discuss the implications of these findings and outline avenues for further research.
Chapter 2

Related Work

This work takes, as a starting point, a large body of existing research on skin photoaging, skin aging, and the assessment of these pathologies using various NGS methods, including scRNA-seq, scATAC-seq, and spatial transcriptomics. It also draws on a range of computational methods that cleverly interrelate these often distinct modalities into a unified whole, allowing for greater insight into the underlying cellular processes that occur at both the genetic and epigenetic levels. This chapter aims to situate this work within the relevant literature.

Skin photoaging is a cumulative process that stems from exposure to UV radiation. Some substantial histological changes associated with photoaging (as well as chronological aging) include reductions in epidermis thickness, decreases in dermal-epidermal junction convolution, and reductions in collagen density in the dermis [43]. Skin photoaging tends to occur in habitually sun-exposed areas of the body such as the face, neck, and arms [21]. Clinically, these changes manifest as wrinkles, lentigines, telangiectasia, mottled pigmentation, roughened texture, sallow complexion, laxity, and decreased elasticity [21]. Physical methods to measure the extent of photoaging include visual assessments, histopathology, and skin echogenicity. Skin echogenicity, discovered by Gniadecka and Jemec, can measure photoaging via decreases in echogenicity in the upper dermis [16]. Moreover, this method can distinguish photoaging from chronological aging, which is associated with an increase in echogenicity in the lower dermis [16].
Photoaging-specific changes are known to stem from the effect UV irradiation has on protein kinase-mediated signaling pathways, such as the upregulation of matrix-degrading genes from the metalloproteinase family (e.g., MMP-1, MM-3), which can decrease the amount of extracellular matrix, thereby impairing the structural integrity of the dermis [11]. Lin et al., among others, showed, using scRNA-seq analysis, that UV radiation escalated the expression of inflammatory factors, especially MMP3, CCL8, and IL [22]. Further, they observed large increases in macrophage cell proportions after UV irradiation, and corresponding increases in the gene HMOX1 [22]. This gene, heme oxygenase, plays a crucial role in anti-inflammation and is expressed by macrophages. Additionally, UVB can interact with intracellular chromophores and photosensitizers, enabling base mutations in DNA molecules [22]. Loss-of-function mutations of TP53, NOTCH1, CASP8, and CDKN2A are associated with cutaneous squamous cell carcinoma and basal cell carcinoma [22].

Similar changes can accompany chronological aging as well. Using scRNA-seq, Zou et al. compared chronological aging between cells from young and old tissue samples, finding crucial differences in the expression of transcription factors (TF) HES1 in fibroblasts and KLF6 in keratinocytes. Inhibition in the expression of these TFs in aged skin tissue contributed to decreased cell proliferation and increased inflammation and cellular senescence [43]. Additionally, they discovered that fibroblasts, endothelial cells, and pericytes are more sensitive to age-related stressors than other cell types [43]. In a similar vein, Sole-Boldo et al. utilized scRNA-seq to discover that aging leads to a loss in dermal fibroblast priming – that is, these cells had a significant delay at the G1/S transition, reduced proliferative capacities, and age-dependent loss of gene expression involved in functional pathways (e.g., collagen secretion) [32]. In this way, different fibroblast lineages tended to become less differentiated over time. Sole-Boldo et al. also observed an upregulation of genes related to immune response and inflammation, and a general decrease in fibroblast cell-cell interactions in the older cells relative to younger cells [32]. Moreover, recent evidence seems to suggest that skin
aging is enforced and accelerated by the accumulation of senescent dermal fibroblasts, which can release proinflammatory, tissue-degrading senescent-associated secretory phenotype factors across the organ system [39].

While great leaps in our understanding of genetic components involved in skin aging and skin photoaging have been made using scRNA-seq techniques, similar studies leveraging scATAC-seq are currently limited. However, scATAC-seq techniques have been insightful in other pathologies, such as systemic sclerosis. Here, Liu et al. used scATAC-seq to compare the epigenetic landscapes of normal and sclerotized tissue, finding the greatest number of differentially expressed chromatin accessibility peaks in dendritic cells, a type of professional antigen-presenting cells [23]. Using scATAC-seq, they delineated other notable forms of regulome divergence from the normal skin, including in marker genes of T Cells [23]. In the psoriasis domain, Bielecki et al. investigated the ways in which innate lymphoid cells (ILC) are shaped by interleukin-23 (IL-23), and, in turn, drive skin thickening in psoriatic disease [2]. Using scATAC-seq, they tested whether ILC chromatin displayed a capacity for the relevant psoriatic transition, finding that, while many of these cells had open regions of chromatin corresponding to downstream cell fate transitions, these regions became increasingly accessible after induction with IL-23 [2]. Foster et al. leveraged scATAC-seq (in conjunction with scRNA-seq and spatial transcriptomics) to identify the role of distinct fibroblast sub-populations in wound healing in a murine model. Here, they managed to uncover four epigenetically distinct fibroblast subpopulations: mechanofibrotic, activated responder, proliferator, and remodeling [12].

Spatial transcriptomics technologies have also been used to study the spatially localized components of several skin pathologies, although research applying this novel technology to the study of skin photoaging is incredibly limited. Using spatial transcriptomics in conjunction with scRNA-seq analysis, Theochardis et al. showed enrichment of unique fibroblast populations expressing MMP1, MMP3, and MMP11, alongside increased M1 macrophage polarization in diabetic foot
ulcer patients with healing wounds [36]. Conversely, patients with non-healing wounds were found to express greater numbers of M2 macrophage markers near the wound site [36]. Foster et al. also leverage spatial transcriptomics to localize the four distinct wound-healing fibroblasts they discovered with respect to the wound-bed [12]. In comparing psoriatic skin with normal skin, Ding et al. discovered that PI3 and S100A8, two neutrophil marker genes, are highly expressed in psoriasis patients and that neutrophilic cells occupy psoriatic skin at rates far higher than normal skin [8]. This spatial context has also been incredibly useful in studying the etiopathogenesis of keloid tissue. Shim et al. used spatial transcriptomics and scRNA-seq to discover the enrichment of keloid-associated fibroblasts in deeper areas of keloid tissue, located mainly around spots with endothelial transcripts, such as VWF and PECAM1, thus implicating the fibrovascular niche in keloid pathogenesis [31].

Recent work has also extended the spatial paradigm to the analysis of chromatin accessibility. Deng et al. devised a method for spatially resolving chromatin accessibility, applying their method to tonsil tissues [7]. In a later work, Deng et al. devised a method for spatial genome and epigenome co-resolution, allowing for a fuller spatial-omics picture in the characterization of a mouse embryo.

As insightful as spatial transcriptomics technologies can be, cost, labor, and skill are three limiting factors that impede their use at scale. To bridge this gap, several spatial transcriptomics inference methods have been developed [9][41][28]. At a high level, these methods decompose WSIs into small image patches, which are then used by computer vision models to output spatial RNA predictions. For example, Monjo et al. developed a convolutional neural network model to infer three breast cancer marker genes in human breast cancer tissue samples [28]. They also developed a semi-supervised variant of their original model that used unlabeled histology images to increase the imputation accuracy of consecutive sections, especially in small datasets [28]. Zeng et al., too, trained a model on the HER2-positive breast tumor spatial transcriptomic dataset as well as the human cutaneous squamous cell carcinoma spatial transcriptomics dataset to infer spatially variable
gene expression [41]. Their method, which they termed HisToGene, utilizes a Vision Transformer (ViT) in conjunction with a Graph Neural Network (GNN) to predict patch-level gene expression. Their model is parametrized according to the zero-inflated negative binomial distribution, which they use to model gene expression across the slide [41]. In another anatomical domain, Fatemi et al. experimented with several different deep-learning approaches to predict spatial gene expression across Visium-co-registered colon WSIs. Here, they experimented with traditional CNNs, ViTs, and GNNs, and compared inference performance at different image patch sizes. They found that larger patch sizes perform better than smaller patch sizes and that Inception and GNN models perform better than their ViT counterparts [9]. They also find that different neural architectures may better predict genes involved in different biological pathways [9].
Chapter 3

Methodology

Section 3.1  
Data Collection

Facial skin tissue was collected from sites histologically adjacent to basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) during skin cancer removal surgery. These tissues were then processed and preserved as either frozen samples or formalin-fixed paraffin-embedded (FFPE) blocks – i.e., permanent samples. In total, 261 skin samples were collected, of which a limited number were used for further genomic analyses. These samples were also given solar elastosis grades by expert dermatologists. In this study, solar elastosis grades were used as a proxy for skin photoaging severity. Solar elastosis grades were classified into three distinct categories: mild, moderate, and severe.

Section 3.2  
Dataset Description

3.2.1. Image Data

Every skin sample, irrespective of processing type, was scanned using the Leica Aperio GT450s developed by Leica Biosystems (Wetzlar, Germany). Each sample, once scanned, ranged in size
Figure 3.1: Hematoxylin and Eosin Stained Facial Skin Image. (A) Hematoxylin stains cell nuclei a deep blue-purple color, while eosin stains proteins like collagen a light pink [10]. (B) The full-size image is so large that individual cells are visible when looking at a 256 by 256-pixel slice of the larger image.

from dimensions as small as 10,000 by 10,000 pixels to images as large as 90,000 by 90,000 pixels. These images were saved as either .tif or .svs files. An example whole slide image is displayed in Figure 3.1.

3.2.2. Genomics Data

Once tissue samples were processed, a subset of these samples were subjected to further genomic analyses. Four distinct analyses were performed on these samples to quantify single-cell and tissue-wide gene expression.

**Single-cell RNA Sequencing.**

**Overview:** ScRNA-seq was used to understand the genes expressed in each cell within a given tissue. This technology has many uses: it can reveal complex and rare cell populations, uncover regulatory relationships between genes, and track the trajectories of distinct cell lineages in develop-
3.2 Dataset Description

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Solar Elastosis Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>F</td>
<td>84</td>
<td>Severe</td>
</tr>
<tr>
<td>167</td>
<td>F</td>
<td>55</td>
<td>Severe</td>
</tr>
<tr>
<td>178</td>
<td>F</td>
<td>29</td>
<td>Mild</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>76</td>
<td>Mild</td>
</tr>
<tr>
<td>261</td>
<td>F</td>
<td>41</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

Table 3.1: ScRNA-seq and ScATAC-seq Data Demographics. Note that all samples were taken from female patients.

In this research, scRNA-seq was used to characterize cell populations and elucidate differentially expressed genes. Once characterized, the effects of solar elastosis were measured by comparing cell type frequencies between different samples. Additionally, the genetic bases of these changes were determined by viewing differential gene expression.

**Data Description:** The Chromium Single Cell Gene Expression assay by 10X Genomics (Pleasanton, CA) was used to measure single-cell gene expression. After initial pre-processing, a gene expression profile was produced for each cell in the tissue sample. Five tissue samples were analyzed using scRNA-seq (solar elastosis grades: two mild, two severe, and one moderate). Of these five samples, two samples had been processed using both a standard tissue freezing workflow, as well as a formalin fixing workflow (i.e., frozen vs. permanent). ScRNA-seq results were encoded in matrix $M$ with dimensions $n_{\text{cells}} \times m_{\text{genes}}$, where $n_{\text{cells}}$ is the number of single cells and $m_{\text{genes}}$ is the number of genes, such that $M_{n,m} \geq 0$ is the expression level of gene $m$ in cell $n$.

**Single-cell Multiome Sequencing.**

**Overview:** Single-cell multiome sequencing includes both scRNA-seq and scATAC-seq assays, uncovering both the genomic and epigenomic landscape within a given set of cells by measuring gene expression and chromatin accessibility. In scATAC-seq, Tn5 transposase integrates into regions of accessible chromatin, whereas steric hindrance in less accessible chromatin makes such transposition less probable [5]. In this way, scATAC-seq profiles the cell epigenome. The multiome assay returns both scRNA-seq and scATAC-seq data for each input cell, allowing for robust
comparisons between the cell genome and epigenome.

**Data Description:** The *Chromium Single Cell Multiome ATAC + Gene Expression* assay by 10X Genomics (Pleasanton, CA) was used to perform both scATAC-seq and scRNA-seq analysis. Five samples were analyzed (solar elastosis grades: two mild, two severe, and one moderate). ScATAC-seq results were encoded in binary matrix $M$ with dimensions $n_{\text{cells}} \times m_{\text{peaks}}$, where $n_{\text{cells}}$ is the number of single cells and $m_{\text{regions}}$ is the number of genome regions evaluated, such that $M_{n,m} = 0$ if genomic region $m$ is not accessible in cell $n$, and $M_{n,m} = 1$ otherwise. Each genomic region activity peak corresponds to a known gene. It is important to note that there is not necessarily a bijection between peaks and genes; some genes may correspond to multiple peaks (e.g., enhancer and promoter activity), and some peaks may regulate several genes.

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Solar Elastosis Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>F</td>
<td>84</td>
<td>Severe</td>
</tr>
<tr>
<td>167</td>
<td>F</td>
<td>55</td>
<td>Severe</td>
</tr>
<tr>
<td>178</td>
<td>F</td>
<td>29</td>
<td>Mild</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>76</td>
<td>Mild</td>
</tr>
</tbody>
</table>

Table 3.2: Visium Spatial Transcriptomics Data Demographics. Four rather than five samples were assayed using Visium spatial transcriptomics (261 was excluded).

**Visium Spatial Transcriptomics.**

**Overview:** The *Visium Spatial Gene Expression* assay by 10X Genomics (Pleasanton, CA) spatially localized gene expression across a WSI, allowing for a more nuanced analysis of gene expression sensitive to spatial variance. This was essential to capture the cellular and spatial heterogeneity of tissue samples. Spatial transcriptomics methods were further paired with scRNA-seq and scATAC-seq methods, allowing for the localization of scRNA-seq data and scATAC-seq data across the WSI.

**Data Description:** Four Visium data samples were collected (solar elastosis grades: 2 mild and 2 severe). The Visium capture area has around 5,000 gene expression spots, where each spot is
55 microns, capturing 1–10 cells per spot [9]. Spatial gene expression data were then aligned, using 10X Genomics’s CytAssist tool (Pleasanton, CA), with a higher definition 40x resolution image. Visium data mirrors the format of both scATAC-seq and scRNA-seq data. Visium results were encoded in a matrix $M$ with dimensions $n_{\text{spots}} \times m_{\text{genes}}$, where $n_{\text{spots}}$ is the number of distinct Visium probes and $m_{\text{genes}}$ is the number of genes measured, such that $M_{n,m} \geq 0$ is the expression level of gene $m$ at spot $n$. An example arrangement of these spots is shown in Figure 3.2.

### Section 3.3

**Tools and Libraries**

#### 3.3.1. ScVI

As powerful as scRNA-seq technologies are, the interpretation of scRNA-seq data remains challenging, as it can be confounded by limited and variable sensitivity, batch effects, and transcriptional noise [25]. To address these challenges, a range of computational models have been developed. One of the most successful approaches is single-cell variational inference (scVI). ScVI uses deep learning to correct these confounds by aggregating information across similar cells and genes and to approximate the distributions that underlie observed expression values [25]. Stated differently, scVI can generate latent embeddings resistant to batch effects and noise for each cell in a scRNA-seq dataset. In this work, ScVI was used to produce latent embeddings for cells in scRNA-seq datasets and perform differential gene expression analysis.

#### 3.3.2. Scanpy

Scanpy is a highly efficient package for analyzing single-cell gene expression data [40]. It includes methods for preprocessing, visualization, clustering, cell development trajectory inference, and differential expression testing, among other important functionalities. In Scanpy, data is stored in the AnnData format, which contains annotations of observations (samples, cells) and variables (features, genes), as well as unstructured annotations. In this work, Scanpy was used to handle all scRNA-seq and Visium spatial transcriptomics data.
3.3 Tools and Libraries

**Methodology**

Figure 3.2: Visium Spatial Transcriptomics Example. (A) Visium sequencing spots displayed across the WSI H&E. Note that every spot corresponds to a spatial capture area of 55 microns. The total slide dimensions are no larger than 6.5 x 6.5 mm. (B) Visium spots colored by their gene expression cluster (as measured by Leiden clustering). (C) Log expression of MLANA, a marker gene for melanocytes (melanin-producing cells), at each Visium spot. (D) Log expression of ADIPOQ, a marker gene for adipocytes (fat-containing cells), at each Visium spot.
3.3.3. Muon

Muon is a multimodal NGS analysis framework that allows easy manipulation of multi-omics data [4]. It extends many of the core Scanpy functionalities beyond scRNA-seq to technologies like scATAC-seq. Here, Muon was used to process scATAC-seq data.

3.3.4. Tangram

Tangram is a method that aligns scRNA-seq data to various forms of spatial transcriptomics data, including Visium spatial transcriptomics data [1]. Tangram aims to learn the alignment between single-cell and spatial transcriptomics datasets by minimizing the Kullback–Leibler divergence between these datasets [1]. In this work, a modified Tangram mapping procedure was devised and implemented to spatially align scRNA-seq and scATAC-seq data with Visium spatial transcriptomics data.

3.3.5. PyTorch

All models used were designed and implemented in PyTorch, a machine-learning library built on the Torch library. PyTorch is highly usable and incredibly fast [29].

### Section 3.4

#### Computational Analyses

3.4.1. Single Cell RNA Analysis

ScRNA-seq data was used to profile gene expression in skin tissue, delineate the effects of solar elastosis on skin tissue cell composition, and investigate differential gene expression between tissues with distinct solar elastosis grades.

**Pre-processing and Integration:** ScRNA-seq data were pre-processed and integrated using Scanpy and scVI, to produce latent vectors resistant to batch effects and other noise. ScVI-based doublet prediction was used to filter out doublet cells (i.e., sequencing errors that combine the gene ex-
expressions of two different cells). Genes not expressed in at least 10 cells and cells expressing less than 200 genes were filtered out. Cells expressing over 10 percent mitochondrial DNA were also filtered out. Gene expression was then log-normalized, and the top 2000 highly variable genes were analyzed.

**Clustering:** Cell-specific latent embeddings were produced for each cell using scVI. These embeddings were used during the application of UMAP [27], a dimensionality reduction technique, which mapped these high-dimensional latent embeddings into two dimensions. The Leiden clustering algorithm [38] was used to cluster UMAP embeddings with a resolution set to 0.3.

**Cell Identification:** Gene markers were used in tandem with label transfer methods to locate distinct cell populations in scRNA-seq data [32][18]. Specifically, a random forest classifier was trained to extend known cell labels from the *Tabula Spaiens Skin Dataset* to unlabeled cells in the collected scRNA-seq datasets. This classifier utilized scVI latent vectors to identify patterns and classify cells based on their unique gene expression profiles. Marker gene-based approaches were layered on top of label predictions in cases where cell type labels were ambiguous. The marker genes utilized are listed in Table 3.3.

**Cell Population Comparisons:** ScVI differential expression was used to compare cell population frequencies across samples stratified by solar elastosis grade.

**Processing Workflow Comparison:** A range of questions exist about how tissue samples should be preserved prior to genome sequencing. Here, the efficacy of two preservation methods – tissue freezing and FFPE – were compared. ScRNA-seq data for both frozen and permanent samples were pre-processed, integrated, clustered, and identified in the same manner described above. Differential gene expression analyses were performed to assess differences in processing workflows.
### 3.4 Computational Analyses

#### Methodology

#### Table 3.3: Cell Types and Marker Genes

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Marker Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macrophages/ Dendritic Cells</td>
<td>LYZ, AI1F1, HLA-DRA, CD68, ITGAX</td>
</tr>
<tr>
<td>T Cells</td>
<td>CD3D, CD3G, CD3E, LCK</td>
</tr>
<tr>
<td>Vascular EC (Endothelial Cells)</td>
<td>SELE, CLDN5, VWF, CDH5</td>
</tr>
<tr>
<td>Lymphatic EC</td>
<td>CLDN5, LYVE1, PROX1</td>
</tr>
<tr>
<td>Pericyte</td>
<td>ACTA2, RGS5, PDGFRB</td>
</tr>
<tr>
<td>Erythrocytes</td>
<td>HBA1, HBA2, HBB</td>
</tr>
<tr>
<td>Melanocytes</td>
<td>PMEL, MLANA, TYRP1, DCT</td>
</tr>
<tr>
<td>Keratinocytes (differentiated)</td>
<td>KRT5, KRT14, TP63, ITGB1, ITGA6</td>
</tr>
<tr>
<td>Keratinocytes (undifferentiated)</td>
<td>KRT1, KRT10, SBSN, KRTDAP</td>
</tr>
<tr>
<td>Fibroblasts</td>
<td>LUM, DCN, VIM, PDGFRA, COL1A2</td>
</tr>
</tbody>
</table>

These genes, also used by Sole-boldo et al., guided cell type assignments in scRNA-seq data [32].

#### 3.4.2. Single Cell ATAC Analysis

**Pre-processing and Integration:** ScATAC-seq data were pre-processed and integrated using Muon. Quality control metrics were first generated for scATAC-seq data, and cells with less than 200 total counts and genes expressed in less than 200 cells were filtered out. Cells were also filtered by nucleosome signal (high signals imply strong sequencing results, low signals imply higher noise), and transcription start site (TSS) enrichment score. TF-IDF and log-normalization were performed on the resulting data.

**Clustering:** Cell-specific embeddings were produced via TF-IDF and latent semantic indexing (LSI). Here, genes in a cell were treated like words and cells were treated like documents. The processed embeddings were clustered using the Leiden algorithm with a resolution set to 0.5. TF-IDF embeddings were reduced to two dimensions using UMAP and displayed.

**Cell Identification:** Due to the limited label transfer methods and greater data noisiness for scATAC-seq data, manual cell identification was performed where possible using a marker-gene-based approach. In cases where concurrent cell data was available via scRNA-seq, labels were transferred using shared single-cell identifiers.
3.4.3. Visium Spatial Gene Expression Analysis

**Pre-processing:** Visium spots and genes were filtered out based on the following criteria: expression of fewer than 500 genes per spot, greater than 15% mitochondrial gene expression per spot, genes were expressed in < 3 distinct cells. Gene expression data was then log-normalized, and the top 2000 highly variable genes were determined.

**Clustering:** UMAP and Leiden were leveraged for dimensionality reduction and clustering. A resolution of 0.5 was used.

3.4.4. Single Cell Multiome Analysis

Four samples possessed all three assay types. Thus, the genetic, epigenetic, and spatial assays were analyzed and compared across samples.

**Pre-processing and Integration:** Samples of each kind were pre-processed as described in section 3.4.2, section 3.4.1, and section 3.4.3.

**Clustering:** Samples were clustered as described in section 3.4.2, section 3.4.1, and section 3.4.3.

**Cell Identification:** Cells in the scRNA-seq data were identified using the same protocol described in 3.4.1. After cell identification, labels were transferred from each cell in the scRNA-seq data to each cell in the scATAC-seq data using their shared identifiers.

**Cell Spatial Mapping:** Tangram was utilized to localize single cells to Visium spots. ScRNA-seq was used to perform the initial correspondence. Due to shared cell identifiers between scRNA-seq and scATAC-seq data, scATAC-seq data was also spatially localized.

**Comparisons:** Leiden clustering results were compared between scRNA-seq, scATAC-seq, and Visium spatial transcriptomics methods. Spatial cell type and cluster localization differences were
3.4 COMPUTATIONAL ANALYSES

3.4.5. Virtual Spatial Gene Expression

Models were trained to infer spatial gene expression and chromatin accessibility from H&E WSIs. To evaluate the functional performance of the spatial RNA expression model, a separate histology classification task was devised.

**Spatial RNA Expression.**

**Pre-processing:** Visum spots and genes were filtered out based on the criteria described in section 3.4.3. The remaining spots were mapped to their corresponding pixel locations on the 40X resolution WSI using CytAssist. Spatial gene expression profiles contained information for upwards of 18,000 genes at 1000s of locations per slide, sampled in a honeycomb formation. Each Visium spot covered a circular capture area of 55 microns, corresponding to an image region of around 260 pixels at 40X magnification, an area large enough to encompass 1-10 cells.

**Data Production:** Square image patches of size $512 \times 512$ pixels were produced such that each patch was centered on a Visium spot. In total, 9388 train and 1737 test patches were created. The SpatialDE package was used to determine the 1000 most spatially varied genes across all four samples (i.e., determined for each slide, and then averaged across all four samples) [33]. Separate models were trained and tested on their ability to recover expression for all 1000 genes for both dichotomized expression (binary classification) and log expression (regression), in line with the procedure shown in Fatemi et al. [9].

<table>
<thead>
<tr>
<th>Name</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-14</td>
<td>Samples 107,167, and 178</td>
<td>Sample 14</td>
</tr>
<tr>
<td>Model-107</td>
<td>Samples 14, 167, and 178</td>
<td>Sample 107</td>
</tr>
<tr>
<td>Model-167</td>
<td>Samples 14, 107, and 178</td>
<td>Sample 167</td>
</tr>
<tr>
<td>Model-178</td>
<td>Samples 14, 107, and 167</td>
<td>Sample 178</td>
</tr>
</tbody>
</table>

**Table 3.4: Spatial RNA Prediction Training Breakdown.** 4-fold cross-validation was used to produce robust performance metrics. The training/testing distribution is shown for each model.
Architecture and Implementation Details: The SWIN-T Vision Transformer model was used to infer spatial gene expression in both the dichotomized and continuous (log-expression) tasks [24]. Images were resized from $512 \times 512$ to $448 \times 448$ pixels. Feed-forward layers of sizes 2000 and 1000 were appended to all models after the final flattening layer. The Lion optimizer was used with task-specific learning rates detailed in the results section (i.e., different between dichotomized and continuous expression tasks) [6]. Models were trained using 4-fold cross-validation and reported performance metrics were averaged across all models. The training setup is shown in Table 3.4.

Spatial ATAC Expression.

Pre-processing: ScRNA-seq data was processed and spatially resolved to corresponding Visium spots using a modified Tangram mapping procedure. In this procedure, Visium data was treated as if it were the scRNA-seq dataset, and the scRNA-seq dataset was treated as if it were the Visium data, allowing for the production of a spot-to-cell matrix $M_{inv}$ (with dimensions $n_{spots} \times m_{cells}$), rather than the conventional cell-to-spot matrix (with dimensions $m_{cells} \times n_{spots}$). Producing the spot-to-cell matrix allowed multiple cells to map to the same spot $s$, such that $\sum_{c \in cells} M_{s,c} = 1$. After obtaining this matrix, scATAC-seq data was spatially localized across the Visium slide using the correspondence between scRNA-seq and scATAC-seq datasets provided via the multiome assay.

The Visium spatial transcriptomics, scRNA-seq, and scATAC-seq data were filtered such that only those genes common to all three assays remained. Then, each Visium spot was assigned a chromatin accessibility vector using probabilities obtained from $M_{inv}$ in conjunction with the scATAC-seq dataset. Specifically, each row, $p_s$, from the inverse Tangram spot-to-cell matrix $M_{inv}$ was taken to give the probability of finding some cell $c$ at some Visium spot $s$. The cell-by-chromatin-accessibility matrix, $M_A$ (with dimensions $n_{cells} \times m_{regions}$, where regions are identical to genes), that described the scATAC-seq data, was then coupled with $p_s$ to produce spot-level accessibility.
Specifically, the chromatin accessibility vector $v_s$ for some spot $s$ was given by the following equation: $v_s = p_s \cdot M_A$. The accessibility vector $v_s$ for some spot $s$ was computed using the weighted average of accessibility vectors found in the scATAC-seq dataset, weighted according to $M_{inv}$.

Rather than incorporating information about the chromatin accessibility of all cells in the production of $v_s$ for some spot $s$, the top 5 cells (as given by their probabilities of being found in spot $s$) were used the weighted average, based on the idea that each Visium spot should only contain between 1-10 cells. The weights for these five cells were adjusted such that they summed to 1. This procedure was repeated for all spots in each Visium slide, producing a chromatin accessibility matrix in place of the original gene expression matrix.

**Data Production:** Square image patches of size $512 \times 512$ pixels were produced such that each patch was centered on a Visium spot. In total, 9696 training and 1849 testing patches were created. The SpatialDE package was applied to find the top 100 most spatially varied genomic regions in terms of chromatin accessibility across all four samples (i.e., determined for each slide, and then averaged across all four samples) [33]. A model was trained and tested on its ability to recover chromatin accessibility for 100 genes in a dichotomized fashion (binary classification), in line with the procedure shown in Fatemi et al. [9].

**Architecture and Implementation Details:** The SWIN-T Vision Transformer model was used to infer spatial gene expression [24]. Images were resized from $512 \times 512$ to $448 \times 448$ pixels. Feed-forward layers of sizes 768 and 100 were appended to the model after the final flattening layer. The Lion optimizer was used with a learning rate set to $5 \times 10^{-6}$ [6]. A batch size of 64 was used. As this model was devised to serve as a proof of concept, 4-fold cross-validation was not performed.
3.4 Computational Analyses

**Methodology**

**Figure 3.3: Example Histological Annotations.** Each dot corresponds to a spatial region with the indicated histology.

**Histology Classification.** To measure the functional performance of the spatial RNA prediction models, simple neural networks were trained using both ground truth gene expression data and synthetic gene expression data to perform histological classification. In this task, given a gene expression vector for some Visium spot, the model is tasked with predicting the spot’s corresponding histological annotation. By training and evaluating these models using both ground truth and synthetic data, the functional quality of the synthetic data was evaluated.

**Pre-processing:** Visium spots were annotated by expert dermatologists in accordance with their histology. These spots were classified as either sebaceous gland, epidermis, hair follicle, eccrine gland, lymphatic/vascular, or nerve (Figure 3.3).

**Data Production:** Gene expression vectors were collected for each spot using both ground truth and predicted data across both the dichotomized and continuous regression tasks.
Architecture and Implementation Details: A three-layer neural network was developed with layer sizes 1000, 500, 50 and an output layer corresponding to each of the 6 target classes. Stochastic gradient descent was used with a learning rate of 0.001 and a momentum of 0.9. The cross-entropy loss was used, and the model was trained with a batch size of 64.
Section 4.1

**Single Cell RNA Sequencing**

**ScRNA-seq analysis of human skin:** After initial preprocessing, where sample cells and genes were excluded as described in section 3.4.1 (i.e., genes < 20 cells, cells < 200 genes), a diverse population of skin cells was integrated across all five samples. The integrated data contained a total of 7042 cells that passed quality controls. The UMAP plot of the integrated data is visible in Figure 4.1, containing 23 clusters, with each cluster representing a distinct cell sub-type. Samples were evenly distributed throughout the integrated dataset. All but one cluster contained data from every sample and represented all three solar elastosis grades. A combination of label transfer from reference datasets and marker-gene-based identification approaches was then used to assign cell identities to each cluster. The reference and query datasets showed great genetic congruence, allowing for successful label transfer (Figure 4.1). The results of this hybrid approach are shown in Figure 4.2.

**Solar elastosis alters endemic cell populations and gene expression:** The effect of solar elastosis severity on cell type frequency was further investigated. Moderate-grade solar elastosis samples were omitted from this analysis. The results of this analysis are found in Figure 4.3. Alterations in
the frequency of various immune cells (CD4/CD8 T cells, macrophages, dendritic cells, and others) as well as in connective tissue (muscle cells, epithelial cells, stromal cells) were found. In samples with severe elastosis, increases in the frequency of CD4 T cells, epithelial cells, macrophages, mast cells, and stromal cells were observed. In samples with mild severe elastosis, increases in CD8 T cells, dendritic cells, endothelial cells, melanocytes, muscle cells, NKT cells, and plasma cells were observed. Note that these differences were not tested for statistical significance.

Genes were also tested for statistically significant differential expression between solar elastosis grades (Figure 4.4), with log-fold change and p-value thresholds set to 3.0 and 0.05, respectively. In cells from samples with severe solar elastosis, IGHG4, IGHG3, AC093001.1, SCGB1B2P, PIP, and IGKC were found to be up-regulated, while NRAP and ASB5 were down-regulated. In cells from samples with mild solar elastosis, NRAP was up-regulated, while IGHG3, IGHG4, AC093001.1, IGHGP, SCGB1B2P, and IGKC were down-regulated. HSPA1A was also found to be up-regulated in cells from samples with mild solar elastosis.

**Processing Modality Affects Resident Cell Populations and Gene Expression.** The frozen and permanent tissue processing workflows were also compared to assess their effect on downstream sequencing. Differential gene expression heatmaps were created for the integrated and labeled scRNA-seq datasets produced from both workflows (Figure 4.5). Cell-specific differentially expressed genes were distinct between the samples in each processing workflow. For all but one cell type (muscle cells), the top 3 cell-specific differentially expressed genes were entirely disjoint between the two processing workflows (Figure 4.6).

---

**Section 4.2 Single Cell Multiome**

Sc-RNA-seq, sc-ATAC-seq, and Visium spatial transcriptomics data were successfully processed and integrated. For each sample, scRNA and scATAC Leiden clusters were determined, and these
4.2 SINGLE CELL MULTIOME

**Figure 4.1:** Integrated scRNA-seq UMAP. (A) Sample ID distribution. (B) Solar elastosis grade distribution. (C) Leiden clustering based on cell RNA expression. Every dot in this figure represents one cell.

**Figure 4.2:** Integrated scRNA-seq Cell Type Assignments. UMAP projection of integrated scRNA-seq data. (A) Cells are colored by sample membership. (B) Cells are colored by the solar elastosis grade of their corresponding samples. (C) Cells are colored according to their Leiden clusters (resolution = 0.5).
4.2 Single Cell Multiome Results

Figure 4.3: Cell Type Frequency by Solar Elastosis Grade. Moderate samples were removed from this analysis. Two samples from each solar elastosis grade were included in this analysis.

Figure 4.4: Differentially Expressed Genes by Solar Elastosis Grade. Volcano plot comparing gene expression in mild and severe cases. Down-regulated genes are colored blue, while up-regulated genes are colored red. (A) In samples with mild solar elastosis, NRAP is up-regulated, while IGHG3, IGHG4, AC093001.1, IGHGP, SCGB1B2P, and IGKC are down-regulated (relative to samples with severe solar elastosis). (B) In samples with severe solar elastosis, NRAP and ASB5 are down-regulated, while IGHG4, IGHG3, AC093001.1, SCGB1B2P, PIP, and IGKC are up-regulated.
### 4.2 Single Cell Multiome Results

**Figure 4.5: Differential Gene Expression Heatmap.** (A) Heatmap created from integrated frozen samples. (B) Heatmap created from integrated permanent samples.
4.2 SINGLE CELL MULTIOME

Figure 4.6: Differentially Expressed Genes by Cell Type. Each subplot shows the top 3 differentially expressed genes for a given cell type by sequencing method (blue and orange indicate permanent and frozen processing workflows, respectively).
cluster labels were transferred between datasets. Cluster labels were also transferred from the single-cell datasets to the Visium samples using the inverse Tangram spot-to-cell mapping matrix. The results of these analyses are shown in Figures 4.7, 4.8, 4.9, 4.10. Cell type identity assignments were performed using scVI label transfer from the reference dataset. These cell-type assignments are visualized for both the scRNA-seq and scATAC-seq datasets, shown in Figure 4.8 and Figure 4.7, respectively. This general processing workflow was applied to all five samples (however, sample 261 did not contain corresponding Visium spatial transcriptomics data). Leiden clustering was used to determine clusters in both scRNA-seq and scATAC-seq data. The correspondence between samples in each dataset was used to freely transfer Leiden labels across assay type (Figure 4.9). A similar analysis was performed using spatial transcriptomics information. The identity of each Visium spot was determined by the identity of the cells most likely to belong to the relevant spot (determined by the inverse Tangram mapping). Leiden cell labels and cell type information were then mapped across the Visium sample (Figure 4.10).

4.3. Virtual Spatial Gene Expression

Models were developed to recapitulate spatial gene expression and spatial chromatin accessibility for highly spatially variable genes using histomorphological data.

4.3.1. Spatial RNA Expression Prediction

**Dichotomized Expression:** The dichotomized prediction tasks consisted of predicting whether gene expression within a defined sample region was higher or lower than the mean expression of that gene across the corresponding WSI. As this task simplified to a discrete binary classification task, AUC-ROC and F1 scores were leveraged to evaluate model performance. MSE loss was also used. The complete dataset consisted of 4 distinct Visium samples, so a 4-fold cross-validation approach was used, where models were trained using three samples and tested on a distinct left-out sample. In this way, four distinct models were trained. The averaged results of model training are
Figure 4.7: scRNA-seq-derived cell types mapped onto scATAC-seq data. (A) Included above are sample 14, (B) sample 261, (C) sample 107, (D) sample 178, (E) sample 167.
Figure 4.8: ScRNA-seq-derived cell types mapped onto scRNA-seq data. Included above are (A) sample 14, (B) sample 261, (C) sample 107, (D) sample 178, (E) sample 167.
Figure 4.9: Single Cell Multiome Analysis. (A) RNA Leiden on scRNA-seq UMAP. (B) ATAC Leiden on scRNA-seq UMAP. (C) ATAC Leiden on scATAC-seq UMAP. (D) RNA Leiden on scATAC-seq UMAP.

displayed in Figure 4.12. Here the 25th, 50th, and 75th percentile AUC and MSE loss scores were averaged across all four models and reported by epoch. Hyperparameters were selected through a coarse hyperparameter grid search, where model performance on the test set was measured using different batch sizes, learning rates, and training durations (Figure 4.11). The grid search was performed using model-14. Optimal performance was observed while using a batch size, learning rate, and training length of 64, $0.5 \times 10^{-6}$, and 20 epochs. These hyper-parameters were then used to train all other models. Model losses converged around epoch 15, while model performances (as measured by AUC) stagnated around epoch 10. These models achieved a median averaged AUC and F1 score of 0.80 and 0.61, respectively, on the testing sets (Figure 4.13A). These models achieved their highest average AUC predictions on ADIPOQ, PLIN1, and PKP3, their most middling performance on SOX9, ZNF331, and RDH11, and their lowest performance on ANKRD35, ALAS1, and MIA (Figure 4.13B). The relationship between model AUC and F1 was also plotted ($\rho = 0.70, p = 2.76 \times 10^{-148}$, Figure 4.13B). A visualization of model-14’s performance across these three gene groups is shown in Figure 4.14. To further evaluate the functional performance
4.3 Virtual Spatial Gene Expression

Figure 4.10: Spatial Multiome Analysis. (A) ScRNA-seq Leiden on Visium sample. (B) ScATAC-seq Leiden on Visium sample. (C) Visium Leiden. (D) ScRNA-seq derived cell types spatially resolved to Visium sample.
of trained models, spatial and reduced dimensionality-aligned clustering analyses were performed using both the synthetic and ground truth Visium data (Figure 4.15). Predicted genes were also stratified into deciles according to their test set performance and assessed via pathway analyses (Table 4.1).

<table>
<thead>
<tr>
<th>Decile</th>
<th>Pathway 1</th>
<th>Pathway 2</th>
<th>Pathway 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Proteins with altered expression in ichthyosis</td>
<td>Genes with mutations associated with ichthyosis</td>
<td>Desmosome dysfunction in cardiomyocyte</td>
</tr>
<tr>
<td>8</td>
<td>Dioxin induced chloracne (hypothesis)</td>
<td>Proteins involved in ichthyosis</td>
<td>Corneodesmosomes in atopic dermatitis</td>
</tr>
<tr>
<td>7</td>
<td>Ichthyosis vulgaris</td>
<td>TGF-beta signaling in skin aging</td>
<td>Proteins involved in ichthyosis</td>
</tr>
<tr>
<td>6</td>
<td>Paneth-cell function in Crohn's disease</td>
<td>PNPLA3 in non-alcoholic fatty liver disease</td>
<td>Adipocyte hypertrophy and hyperplasia</td>
</tr>
<tr>
<td>5</td>
<td>Proteins involved in ependymoma</td>
<td>Glioblastoma, proneural subtype</td>
<td>ADORA2A/B → vasodilation</td>
</tr>
<tr>
<td>4</td>
<td>Extracellular matrix turnover</td>
<td>Cathepsins role in periodontitis</td>
<td>Amyloid beta clearance in Alzheimer disease</td>
</tr>
<tr>
<td>3</td>
<td>Metals and amyloid beta toxicity in Alzheimer's disease</td>
<td>Mast-cell activation without degranulation through CXCR4 signaling</td>
<td>Omega-3-fatty acid metabolism</td>
</tr>
<tr>
<td>2</td>
<td>Eicosanoids in inflammation</td>
<td>Dioxin induced chloracne (hypothesis)</td>
<td>Androgen receptor/prostate specific antigen signaling in prostate cancer</td>
</tr>
<tr>
<td>1</td>
<td>Proteins with altered expression in vitiligo</td>
<td>Neuroinflammation in amyotrophic lateral sclerosis</td>
<td>Hemoglobin reduction and leukocyte adhesion initiate vasospasm</td>
</tr>
<tr>
<td>0</td>
<td>IL13R→ STAT6 signaling</td>
<td>IL4R→ STAT signaling</td>
<td>IL21R→ STAT signaling</td>
</tr>
</tbody>
</table>

Table 4.1: Pathways Represented by Gene Performance Decile (Dichotomized). Genes were stratified into deciles by their performance in the dichotomized prediction task. The top 10% of genes were analyzed in decile 9, while the bottom 10% of genes were analyzed in decile 0. Pathway analysis was performed on each performance decile. Displayed are the top 3 pathways for each decile by combined score (i.e., magnitude and statistical significance).

**Continuous Expression:** Models were also trained to recapitulate continuous spatial gene expression. In this task, models inferred log-transformed gene expression from histomorphological data. Similar to the dichotomized expression task, 4-fold cross-validation was used to evaluate model performance. Accordingly, four models were trained, and their performance metrics were collected and averaged. Spearman coefficients were used to measure model performance. MSE
4.3 Virtual Spatial Gene Expression

Results

Figure 4.11: Hyperparameter Grid Search. (A) Batch size, learning rate, and training length (epochs) were systematically altered to assess their impact on test AUC. (B) Learning rate, training length, and batch size were inversely correlated with model performance.

![Hyperparameter Grid Search](image)

Figure 4.12: Dichotomized Expression SWIN-T Model Training. A batch size of 64, a learning rate of $0.5 \times 10^{-6}$, and a training length of 20 epochs were used. (A) Mean 25th, 50th, and 75th, percentile AUC calculated after each training epoch. (B) Mean training loss per epoch. Note that these values are averages from all four trained models.

![Dichotomized Expression SWIN-T Model Training](image)
loss was also used here. Model hyperparameters were selected through a coarse hyperparameter grid search performed using model-14. The results of this analysis underscored the sensitivity of model performance to appropriately low learning rates and small batch sizes (Figure 4.16). A batch size, learning rate, and training length of 64, $0.33 \times 10^{-6}$, and 20 epochs were used across all models. The averaged results of model training are visible in Figure 4.17. Model losses converged around epoch 18, while model performance – as measured by the Spearman coefficient – stagnated around epoch 10. These models achieved a median average Spearman coefficient of 0.60 on the testing sets (Figure 4.18A). Models achieved their highest average performance predictions on KRT14, CXCL14, and COL1A2, their most middling performance on LHFPL6, LCE1A, and MPEG1, and their lowest performance on CKM, MYLPF, and ODF21 (Figure 4.18B). Both the dichotomized and continuous prediction models inferred expression for the same set of 1000 spatially variable genes, and their performance relationship is shown in Figure 4.18B ($\rho = 0.49$, $p = 1.75 \times 10^{-61}$). A visualization of model-14’s performance across these three gene groups is
Figure 4.14: Dichotomized RNA Expression Prediction. Model-14 was used to infer dichotomized spatial gene expression for sample 14 and compared with the ground truth. A spot is colored yellow if gene expression in this spot exceeds global mean gene expression. (A) H&E slide. (B) Top performing genes, from left to right, ADIPOQ, PLIN1, PKP3. (C) Median performing genes, from left to right, SOX9, ZNF331, RDH11. (D) Worst performing genes, from left to right, ANKRD35, MIA, ALAS1.
Figure 4.15: Dichotomized Functional Clustering Analysis. Model-14 was used to infer dichotomized spatial gene expression for sample 14 and compared to the ground truth. (A) Inferred Visium aligned-UMAP. (B) Ground truth Visium aligned-UMAP. (C) Inferred Visium. (D) Ground truth Visium. Spots are colored by Leiden cluster affiliation.
4.3 Virtual Spatial Gene Expression

Results

displayed in Figure 4.19. To further evaluate the functional performance of this approach, spatial and reduced dimensionality-aligned clustering analyses were performed using both the synthetic and ground truth Visium data (Figure 4.20). Predicted genes were also stratified into deciles according to their test set performance and assessed via pathway analyses (Table 4.2).

<table>
<thead>
<tr>
<th>Decile</th>
<th>Pathway 1</th>
<th>Pathway 2</th>
<th>Pathway 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Corneodesmosomes in atopic dermatitis</td>
<td>Desmosome dysfunction in cardiomyocyte</td>
<td>Desmosome assembly</td>
</tr>
<tr>
<td>8</td>
<td>Proteins involved in ichthyosis</td>
<td>Genes with mutations associated with ichthyosis</td>
<td>Lipogenesis/lipolysis activation in cancer cells and cancer-associated adipocytes</td>
</tr>
<tr>
<td>7</td>
<td>Hyperosmotic stress in diabetic neuropathy</td>
<td>Proteins with altered expression in acne vulgaris</td>
<td>Proteins involved in ichthyosis</td>
</tr>
<tr>
<td>6</td>
<td>ADORA2A/B→ Vasodilation</td>
<td>PTGFR→Vasoconstriction</td>
<td>Oxytocin action in mammary gland</td>
</tr>
<tr>
<td>5</td>
<td>Omega-3-fatty acid metabolism</td>
<td>MYOC in cell adhesion in glaucoma</td>
<td>Omega-6-fatty acid metabolism</td>
</tr>
<tr>
<td>4</td>
<td>Adipocyte hypertrophy and hyperplasia</td>
<td>Lipodystrophy, familial partial</td>
<td>Lipogenesis regulation in adipocyte</td>
</tr>
<tr>
<td>3</td>
<td>Proteins with altered expression in vitiligo</td>
<td>T-cells passage through the blood-brain barrier</td>
<td>Proteins involved in dental caries</td>
</tr>
<tr>
<td>2</td>
<td>Eicosanoids in inflammation</td>
<td>Androgen receptor/prostate-specific antigen signaling in prostate cancer</td>
<td>Proteins involved in sleep initiation and maintenance disorders</td>
</tr>
<tr>
<td>1</td>
<td>Proteins with altered expression in Parkinson’s disease</td>
<td>Thymic follicular hyperplasia</td>
<td>Tumor-infiltrating macrophages in cancer progression and immune escape</td>
</tr>
<tr>
<td>0</td>
<td>Hormones underlie human gastrointestinal functions and eating behavior</td>
<td>N-cell: Neurotensin production</td>
<td>Gluconeogenesis impairment in non-alcoholic fatty liver disease</td>
</tr>
</tbody>
</table>

Table 4.2: Pathways Represented by Gene Performance Decile (Continuous). Genes were stratified into deciles by their performance in the continuous prediction task. The top 10% of genes were analyzed in decile 9, while the bottom 10% of genes were analyzed in decile 0. Pathway analysis was performed on each performance decile. Displayed are the top 3 pathways for each decile by combined score (i.e., magnitude and statistical significance).

Histology Classification: A simple neural network was also trained to evaluate the functional performance of synthetic Visium data in the histological classification task. In this task, the model was trained to infer the histological identity of a Visium spot using both synthetic and ground truth
4.3 Virtual Spatial Gene Expression

**Figure 4.16: Continuous Expression Hyperparameter Grid Search.** Batch size, learning rate, and training length (epochs) were varied to determine their effect on model performance. Lower learning rates and shorter training times resulted in higher model performance.

**Figure 4.17: Log Expression SWIN-T Model Training.** Batch size: 64; learning rate: $10^{-5}$; Models were trained for 20 epochs. (A) Mean 25th, 50th, and 75th percentile performance (as measured by the Spearman coefficient) on the test set per epoch. (B) Mean training loss per epoch. Note that these values were averaged across all four models.
4.3 **VIRTUAL SPATIAL GENE EXPRESSION**

---

**Figure 4.18: Continuous Model Performance.** (A) Mean Spearman coefficients and mean p-values for all tested genes. (B) Mean AUC plotted against mean Spearman coefficient for each gene. Note that both the dichotomized and continuous models were predicting the same set of 1000 spatially variable genes.

Visium gene expression vectors. Models were trained to perform this task using both dichotomized and continuous gene expression vectors. In the dichotomized task, the simple neural network was able to achieve a maximum classification accuracy of 0.93 and 0.87 using ground truth and synthetic Visium data, respectively (Figure 4.21A). In the continuous task, the simple neural network was able to achieve a maximum classification accuracy of 0.92 and 0.88 using ground truth and synthetic Visium data, respectively (Figure 4.21B).

---

**4.3.2. Spatial ATAC Prediction**

**Spatial Chromatin Accessibility Mapping:** Chromatin accessibility data was mapped across the slide as described in section 3.4.5. The general mapping procedure was validated first using scRNA-seq data. Specifically, scRNA-seq data was mapped across the slide and compared to the ground truth Visium sample (Figure 4.22). After the feasibility of this method was confirmed, chromatin accessibility data was mapped across the slide, and the accessibility profile was qualitatively compared to the Visium ground truth (Figure 4.23). Chromatin accessibility was not
Figure 4.19: Continuous RNA Expression Prediction. Model-14 was used to infer log spatial gene expression for sample 14 and compared to the ground truth. This performance is measured in terms of Spearman correlation. (A) H&E slide. (B) Top performing genes, from left to right, KRT14, CXCL14, and COL1A2. (C) Median performing genes, from left to right, LHFPL6, LHFPL6, MPEG1. (D) Worst performing genes, from left to right, CKM, MYLPF, ODF2L.
4.3 Virtual Spatial Gene Expression

Results

Figure 4.20: Continuous Functional Clustering Analysis. Model-14 was used to infer continuous spatial gene expression for sample 14 and compared to the ground truth. (A) Inferred Visium aligned-UMAP. (B) Ground truth Visium aligned-UMAP. (C) Inferred Visium. (D) Ground truth Visium. Spots are colored by Leiden cluster affiliation.
4.3 Virtual Spatial Gene Expression

Figure 4.21: Histology Classification. A simple neural network was trained to predict histology identity from a gene expression vector using (A) dichotomized expression and (B) log expression. In each case, the ground truth vector was compared to its corresponding inferred vector.

more elaborately compared to ground truth spatial gene expression data because these data capture fundamentally distinct -- though related-- facets of the underlying biology (i.e., epigenome vs. genome). Spatial chromatin accessibility maps were produced for all four Visium samples.

Dichotomized Expression: A model was trained to predict dichotomized spatial chromatin accessibility for the top 100 spatially varied genes, determined via SpatialDE, using the data produced by spatial chromatin accessibility mapping. This task consisted of predicting whether chromatin accessibility for some gene in a given spatial region was higher than chromatin accessibility for that gene across the WSI. Similar to dichotomized spatial gene expression prediction, this task simplified to a binary classification problem across spatially variable genes. AUC-ROC and F1-score were leveraged to measure model performance. MSE loss was also used. A batch size, learning rate, and training length of 64, $0.33 \times 10^{-5}$, and 30 epochs were used. Model loss converged within the first 5 epochs. Model test performance also stagnated within the first 5 epochs. 25th, 50th, and 75th percentile test AUCs were 0.74, 0.71, and 0.67, respectively. 25th, 50th, and 75th percentile test F1-scores were 0.66, 0.63, and 0.60, respectively. A few of the best-performing genes achieved a test AUC of 0.80 and F1-score of 0.75 (Figure 4.24).
4.3 Virtual Spatial Gene Expression

Results

Figure 4.22: Spatial Mapping Validation. To validate the synthetic ATAC data produced, the same procedure was used to create synthetic RNA data, which was compared to the ground truth Visium gene expression data. (A) Synthetic dichotomized spatial gene expression. (B) Visium dichotomized spatial gene expression. (C) Synthetic continuous spatial gene expression. (D) Visium continuous spatial gene expression.
4.3 Virtual Spatial Gene Expression Results

**Figure 4.23:** Spatial Chromatin Accessibility Mapping. (A) Synthetic spatial chromatin accessibility of CRAB2. (B) Visium spatial gene expression for CRAB2.

**Figure 4.24:** Dichotomized ATAC Expression Prediction. (A) ATAC prediction for IBA57, an iron-sulfur cluster assembly factor, colored by logit values. (B) Dichotomized ground truth. AUC: 0.80, F1: 0.75.
Chapter 5

Discussion

Existing work on skin photoaging, particularly in humans, is sparse, with the majority of NGS studies investigating skin aging more generally [32][43] [19] [30]. These studies have made significant progress in the identification of genetic changes associated with skin aging (e.g., fibroblast de-differentiation and increases in senescent connective tissue, increased inflammation). However, the study of skin photoaging – those genetic and histological changes in the skin that occur due to UV exposure – is still in its infancy. Using scRNA-seq techniques, this study presents a preliminary analysis of cellular and genetic changes accompanying skin photoaging using over 14648 single cells (7042 cells using the frozen workflow and 7606 cells from the permanent workflow). Furthermore, this study leverages a multi-omics approach to understand skin tissue heterogeneity from a genetic and epigenetic perspective. Lastly, this research proposes and develops novel spatial transcriptomics mapping and prediction methods to democratize access to spatial transcriptomics techniques for future research. Advancements in these techniques will facilitate a deeper understanding of skin photoaging and other skin pathologies at the population level.

5.0.1. ScRNA-seq Analysis:

Many of the major cell types and categories described in existing scRNA-seq work on skin aging were recapitulated in this work[43]. It was discovered, however, that both cell type frequencies and underlying gene expression varied by solar elastosis grade (mild or severe). For instance, IGHG4,
IGHG3, AC093001.1, SCGB1B2P, PIP, and IGKC were up-regulated in cells derived from tissues with severe solar elastosis. Analyzing these genes using the EnrichR analysis platform revealed that these genes were likely to be involved in the complement cascade and innate immune system more generally [20]. On the other hand, NRAP and ASB5 were down-regulated in these same cells. The Reactome database suggests that these genes may be involved in neddylation and ubiquitination and, therefore, part of the broader adaptive immune system [15]. The reverse differential expression pattern occurs in cells acquired from samples with mild solar elastosis. These findings may suggest that photoaging differentially affects immune-related functions in skin tissue: increasing the activity of the innate immune system while decreasing the activity of the adaptive immune system.

These results also indicate that skin photoaging changes the frequency of different cell type populations. General increases in the frequency of some immune cell subpopulations, especially CD4, CD8, regulatory T cells, macrophages, and mast cells, were observed in samples with severe solar elastosis. On the other hand, these same sample populations showed a decline in the frequency of dendritic cells, NKT cells, and plasma cells. These results are partially consistent with the aforementioned gene pathways analysis. Macrophages and mast cells play notable roles in the innate immune system, and their increased frequency of occurrence is consistent with the elevated expression of IGHG4, IGHG3, AC093001.1, SCGB1B2P, PIP, and IGKC. However, the increases in CD4, CD8, and regulatory T cells are harder to compatibilize with the decreased expression of genes related to the adaptive immune system. The increased frequency of these cells may relate to the down-regulation of NRAP and ASB5. If these genes play a crucial role in cell ubiquitination, then their down-regulation may increase the frequency of immune cell receptors and immune cells that would have otherwise been marked for degradation.

Sample processing workflow can also have a substantial impact on downstream scRNA-seq analysis. It was found that the differential gene expression profile of tissue samples varies considerably
by processing type (i.e., frozen or permanent tissue processing and fixing methods). Differential gene expression profiles between frozen and permanent tissue samples were almost entirely distinct. Though the same cell types were present in each sample, the genes most differentially expressed in these cells do not overlap. Only one cell – muscle cells – did not possess entirely distinct cell type differential expression (measured by the top 3 most differentially expressed genes), where CASQ2 was shared between the groups. While more work is needed to elucidate the exact nature of these changes, these results provide initial evidence that tissue processing workflows may substantially impact downstream NGS applications.

### 5.0.2. Multi-omics Analysis:

On the multi-omics front, this work confirms existing knowledge that scATAC-seq and scRNA-seq track very different genetic and regulatory elements. Moreover, scRNA-seq clustering and integration produced clusters that were generally more distinct and easier to characterize than scATAC-seq data. ScRNA-seq cell type labels also cleanly mapped onto scATAC-seq UMAP projections, representing a way to combine scRNA-seq and scATAC-seq data in future multi-omics studies. ScRNA-seq and scATAC-seq Leiden cluster labels were also freely swapped between UMAP projections, allowing visual representation of genomic and epigenomic interplay within samples. These analyses demonstrated how the epigenetic regulatory mechanisms converge and diverge from the underlying genetic expression. This analysis may be particularly useful for understanding how normal skin aging and skin photoaging come apart. Though these two separate pathologies may exhibit similar alterations in gene expression, the driving regulatory forces might look very different.

Furthermore, using the inverse Tangram mapping procedure, single-cell information from both scRNA-seq and scATAC-seq data were spatially mapped across the Visium data. This technique spatially resolved genetic and transcriptional landscapes. Both spatially resolved scRNA-seq and scATAC-seq were sensitive to genome and epigenome information in a way that the standard Vi-
sium slide was not, likely due to their single-cell precision. ScRNA-seq cell type identities were also spatially mapped using the same procedure, where the identity of the spot was determined according to the identity of the cell most likely to be localized there, providing a way of characterizing Visium spot identities in terms of the cells composing the spot. It is worth noting, however, that single-cell data were aggregated across all of these mapping analyses to produce spot-level data. Accordingly, these techniques were not exempt from concerns about limited spatial precision.

5.0.3. Spatial RNA Prediction:

Deep learning may also represent a viable path to overcome current barriers in the adoption of spatial transcriptomics technologies, enabling their use at the population level. This work provides evidence that Vision Transformers can learn to predict spatially variable genes efficiently and accurately across both dichotomized and continuous regression tasks, reaching a median average AUC and F1-score of 0.80 and 0.61 on the dichotomized task and a median average Spearman coefficient of 0.60 on the continuous task.

Beyond numerical performance, it was imperative to understand the extent to which synthetic data captured the essence of ground truth spatial transcriptomics data and whether synthetic data could productively enable downstream analyses, such as histological classification. Accordingly, aligned-UMAP was used to measure the extent to which ground truth and synthetic data agreed in essence (Figures 4.15 and 4.20). While model predictions captured the general shape of ground truth data, they also produced a substantial number of genetically ambiguous spots. In particular, these models tended to make predictions that had intermediate genetic states. This is reflected in both the shape and clustering of the aligned-UMAP plots, where the predicted data tended to form intermediate clusters not present in the ground truth data. Furthermore, spatial clustering analysis showed increased noise present in the predicted data compared to the ground truth data (Figures 4.15 and 4.20, C & D). For example, spots in the tissue hypodermis were associated with more
distinct Leiden clusters in the predicted data than in the ground truth data. This is true for most histological structures, with the exception of the epidermis, which was consistently homogeneous in both the predicted and ground truth data.

The histological classification task was another proxy used to evaluate the quality of inferred spatial transcriptomics data. Results indicated that, while differences exist in model performance dependent on training data (i.e., training on either synthetic or ground truth Visium data), these differences were narrow. Surprisingly, performance gaps were slightly higher between synthetic and ground truth Visium data when the gene expression vectors were dichotomized than when they were continuous. This is especially surprising given that inference in the dichotomized task is theoretically simpler than inference in the continuous regression task. Further, models trained on synthetic data (whether dichotomized or continuous) improved more rapidly than models trained using ground truth data. Models trained on synthetic data also tended to achieve lower end-training performances than those trained on ground truth data. Together, these results suggest that synthetic data is likely noisier, less complex, and more strongly biased than ground truth Visium data which, in turn, allows other models to learn the histological classification task more rapidly using synthetic data, but the shallowness of the data ultimately bottlenecks end-classification accuracy.

Models were also assessed for their performance across individual genes. The top-performing genes in the dichotomized prediction task were ADIPOQ, PLIN1, and PKP3, which function in regulating glucose levels, fatty acid breakdown and lipolysis, and desmosomal binding and cellular adhesion [34] [14] [3]. The worst-performing genes in the dichotomized expression task were ANKRD35, ALAS1, and MIA. ALAS1 and MIA are known to function in heme production and extracellular matrix production, respectively [20] [13], while the role of ANKRD35 is less clear. Disparities in performance between the best and worst-performing genes seem primarily explained by differences in spatial expression frequency. That is, ADIPOQ, PLIN1, and PKP3 were expressed at many locations across the slide, whereas ANKRD35, ALAS1, and MIA were
expressed at only a handful of locations. A model of this size – with nearly 3 billion parameters \[24\] – is likely not large enough to capture the extensive spatial variation in gene expression across 1000 distinct genes. However, future work is needed to clarify the precise relationship between model size, predicted gene set size, and model performance in this task.

The top-performing genes in the continuous prediction task were KRT14, CXCL14, and COL1A2, which are known to function in collagen biosynthesis as well as immune cell migration and antimicrobial immunity [20][15] [26]. The worst-performing genes, by contrast, were MYLPF, ODF21, and CKM, which are thought to play a role in creatine metabolism and the urea cycle [15]. Disparities in performance here, as in the case of the dichotomized task, seem most strongly related to differences in spatial expression frequency. These disparities represent rich areas for further research, as it is presently unknown to what extent genes involved in different biological pathways are systematically better or worse predicted in dichotomized and continuous prediction tasks.

In addition to analyses of select genes, all 1000 genes were stratified into deciles according to their average inference performance. A pathway analysis was then performed on each decile. In the dichotomized task, genes in the top-performing decile were involved in pathways related to ichthyosis and desmosome dysfunction. More generally, genes in the top 3 performance deciles tended to involve pathways implicating ichthyosis, corneodesmosome function, and TGF-β signaling. Genes in the worst performing decile, on the other hand, primarily represented the interleukin-transcription factor signaling pathway (e.g., IL13R → STAT6 signaling), and genes involved in the worst three performance deciles implicated inflammation and androgen receptor signaling, among others.

A similar analysis was conducted on genes predicted in the continuous task. Here, genes in the top-performing decile were involved in pathways related to corneodesmosome and desmosome dynamics, and genes in the top 3 performing deciles were likely to be involved in lipogenesis/lipolysis.
and ichthyosis. Some divergence in associated pathways was observed between the dichotomized and continuous modeling approaches in the worst-performing genes. In the continuous task, genes in the worst-performing decile tended to implicate pathways like gastrointestinal function, gluconeogenesis impairment, and various kinds of neuronal and immune dysregulation, among others.

Together, these pathway analyses illustrated the differential sensitivity of modeling approaches to biological pathways. Specifically, those pathways positioned to alter histomorphology directly tended to be associated with genes in the highest-performing deciles. In contrast, those pathways more distantly causally related to underlying histomorphology tended to be associated with genes in the lowest-performing deciles. In other words, the predictability of a gene seems related – though not wholly determined – by its histomorphological salience. As a concrete example, consider that in the dichotomized task, genes associated with ichthyosis, desmosome and corneodesmosome function, and TGF-β signaling were best predicted, while genes associated with immune signaling were predicted with lower accuracy. In this example, pathways associated with the best-performing genes tended to function directly at the histomorphological level. Ichthyosis pathways involve filaggrin, which binds to intermediate filaments (i.e., keratin) in epithelial cells. Desmosome and corneodesmosome dynamics directly mediate cell-cell adhesion. Finally, TGF-β signaling regulates cellular senescence. All three of these pathways directly regulate either the shape and topology of the cell itself or the spatial proximity of the cell in relation to its neighbors, so, unsurprisingly, this modeling approach – which makes predictions based on histomorphological data – can predict associated genes with high accuracy. Nonetheless, further study is required to understand the strengths and weaknesses of these models, particularly across different biological pathways. These results will inform practical decisions regarding when and where these models can provide insight.
5.0.4. Spatial Chromatin Accessibility Mapping and Prediction:

The final significant contribution of this work was the development of spatial chromatin accessibility mapping and prediction techniques. Spatial chromatin mapping and prediction are feasible tasks, though rigorous validation protocols are needed to ascertain the strength of this mapping. The mapping approach pursued here transferred scATAC-seq data to each Visium spot by first mapping scRNA-seq data to the Visium slide and then using the correspondence between cells in both the scRNA-seq and scATAC-seq datasets to map scATAC data to the Visium slide. To do this, the conventional Tangram mapping procedure – used to map individual cells to Visium spots – was slightly modified. Rather than producing a mapping from cells to spots, the conventional procedure was inverted to produce a mapping from spots to cells. This allowed each spot to be given a probability distribution over all cells in the single-cell dataset. Synthetic gene expression and chromatin accessibility vectors for each spot were then produced by performing a weighted average across the gene expression and chromatin accessibility vectors corresponding to the five cells with the highest likelihood of belonging to the spot. The weighted average was performed using five cells due to the fact that Visium spots cover a surface area large enough to accommodate between 1 and 10 cells. Still, more research is required to determine how many cells need to be incorporated to produce high-quality synthetic spot data. These results must also be heavily caveated because they were not rigorously validated due to the lack of corresponding spatial chromatin accessibility data. Though synthetic spatial chromatin accessibility data was compared to Visium spatial transcriptomics data, these analyses are unlikely to measure the validity of this mapping. To reiterate, if these techniques are to be implemented further, then they must be rigorously tested using spatial epigenome–transcriptome co-profiling techniques, such as those proposed in Zhang et al. [42].

A deeper understanding of skin photoaging is required to prevent the range of malignant skin pathologies with which it is associated. This study has presented a preliminary analysis of the genetic changes related to skin photoaging. It has also developed a range of tools positioned to bring
the spatial transcriptomics paradigm to broader swaths of the research community. Ultimately, I hope the tools, techniques, and approaches proposed in this research will become more than just words on a page.
Chapter 6

Supplementary Materials
Figure 6.1: Dichotomized RNA Expression Prediction. Model-178 was used to infer dichotomized spatial gene expression for sample 178 and compared with the ground truth. A spot is colored yellow if gene expression in this spot exceeds global mean gene expression. (A) H&E slide. (B) Top performing genes, from left to right, ADIPOQ, PLIN1, PKP3. (C) Median performing genes, from left to right, SOX9, ZNF331, RDH11. (D) Worst performing genes, from left to right, ANKRD35, MIA, ALAS1.
Figure 6.2: Continuous RNA Expression Prediction. Model-178 was used to infer log spatial gene expression for sample 178 and compared with the ground truth. This performance is measured in terms of Spearman correlation. (A) H&E slide. (B) Top performing genes, from left to right, KRT14, CXCL14, and COL1A2. (C) Median performing genes, from left to right, LHFPL6, LHFPL6, MPEG1. (D) Worst performing genes, from left to right, CKM, MYLPF, ODF2L.
Bibliography


[24] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo, Swin Transformer V2: Scaling Up Capacity and Resolution, April 2022, arXiv:2111.09883 [cs].


