THE CELLULAR ECOLOGY OF THE MOUSE THYMUS

by

MARY ROUGEAU BROWNING

(Under the Direction of Nancy R. Manley and John P. Wares)

ABSTRACT

The ability to distinguish organizational patterns of cells is essential for understanding organ maintenance and function; however, statistical methods for quantifying cellular organization do not exist. The development of techniques to quantitatively identify recurrent or cryptic cellular patterns could help us to better understand healthy tissue states, and would allow us to make comparisons with tissues in states of disequilibrium. Here, we have developed novel computational analyses that allow us to study the organ of our choice, the thymus, in a purely statistical framework. We accomplish this by borrowing techniques commonly used in ecology and applying them to the thymus; we are able to make this comparison due to the remarkable similarities in behavior and function exhibited by both cell types in organs and species in ecological communities. The output generated from these analyses can be used to better understand healthy thymic homeostasis and states of disequilibrium.

INDEX WORDS: biogeography, quantitative, immunohistochemistry, thymus, K-means, Bray-Curtis, connected component labeling, AIRE.

THE CELLULAR ECOLOGY OF THE MOUSE THYMUS

by

MARY ROUGEAU BROWNING

B.S., Mississippi State University, 2012

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2015

THE CELLULAR ECOLOGY OF THE MOUSE THYMUS

by

MARY ROUGEAU BROWNING

Major Professor:	Nancy Manley, John Wares
Committee:	Andrew Park
	Brian Condie

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
May 2015

DEDICATION

I would like to dedicate this work to my wonderful husband Matthew Browning and my family. To my husband, thank you for your continuous support and guidance and your endless patience and love. To my family, thank you for being with me every step of the way during this journey and providing me with unending encouragement.

ACKNOWLEDGEMENTS

I would like to thank my mentors John Wares and Nancy Manley and my committee members Brian Condie and Andrew Park for their guidance and support during the development and execution of this project.

I would also like to thank my fellow lab members and office staff. To Rodney and Trent, I really appreciate your friendship as well as your help with any questions I had. To Kristina and Julie, thank you for your continuous willingness to help me with anything related to my project. To Jenna, thank you for your wisdom, support, and for serving as a mentor to me during my time as a graduate student; I am very thankful for the friendship we developed. I would also like to thank Joelle, who I consider to be a friend as well as a co-worker.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

QUANTITATIVE ANALYSIS OF CELLULAR ORGANIZATION

Introduction

Within the field of organ biology, there is a lack of quantitative methods for analyzing the organization, as opposed to the composition, of organs and tissues. Healthy organs require more than the appropriate proportions of component parts; these parts must also be arranged in specific ways in order to function properly. Without quantitative methods for assessing organization, a significant aspect of organ biology is inaccessible to analysis. However, beyond essentially descriptive assays, there are few statistical tools available for interpreting organizational characteristics.

In contrast, the fields of ecology and population genetics have a robust mathematical tradition, with multiple quantitative tools available for describing the natural world. Fortunately, and surprisingly, a striking comparison can be drawn between the interactions that occur within organ systems and those that occur within ecosystems. For instance, cell types in an organ, like species in an ecosystem, are influenced by the availability of space and resources as well as by the presence or absence of other cell types. The presence or absence of one cell type may alter the functional role and dominance of another, or the overall service the community provides to the larger system (Wootton, 2005).

Similarly, organs are composed of migrating and resident cell types. Resident cell types can be compared to the vegetation in a community; they maintain the structure of the environment and provide the resources necessary to support the migrating cell types.

Migrating cell types, on the other hand, are like the animals; they travel through the community, interacting with other species while fulfilling their role of both providing and consuming resources. Tissues are not typically composed with all cell types at equal abundance, which is similar to species in a community; there are a few common "species", and many low abundance "species" that fit a lognormal abundance distribution curve (Krebs, 1999).

Finally, both cellular and species populations are considered to have microenvironments and niches with which they have specialized interactions and relationships. In both a biological organ and a natural ecological community, there is a high degree of interaction and interdependence between the system components. Thus, perturbation of one species or cell type can have multiple effects on the entire structure and function of the ecosystem or tissue that can be read out in changes in species abundance or spatial distribution. By comparing cell types within a biological organ to species in an ecosystem, statistical methods used by ecologists can be directly applied to biological systems. We seek to make this comparison in an effort to expand the quantitative methods available to organ biologists through the use of a well characterized and respected field. Our overall goal is the development of a cellular ecology, where we can model cellular relationships and patterns to better understand the "biogeography" of the organ we wish to study. We aim to develop novel quantitative tools, modeled after the ecological toolkit, to spatially and statistically characterize healthy biological tissues. The results from this analysis will enable further informative comparisons with tissues in states of disequilibrium, including disease or developmental states, in order to identify any recurring organizational differences that can potentially lead to dysfunction of the organ of interest.

Our goal is to model a healthy adult organ with the intent of creating a baseline model that can be used for further comparisons. We will accomplish this by identifying location and abundance of different cell types. We will then spatially and compositionally characterize the organ by applying statistical methods commonly used in ecology directly to the produced model. For instance, ecological approaches can describe the degree to which individuals of a species aggregate, and to mathematically describe the size distribution of these aggregations (Hubbell, 2001). By analyzing the exact spatial location and co-location of component cell types across the domain of a particular tissue or organ, we may gain a better understanding of which interactions - indicated by co-distribution and sufficient density - can be validated by independent approach and analysis. To an extent, we recapitulate the spatial scales described above: we explore the differentiation of distinct habitats within the environment (biogeography: identifying regions of endemicity or dramatically shifted abundance), the co-localization of particular sub-groups of cells (community ecology), and we use the tenets of macroecology to the extent that we can generalize that cells need a particular density to be viable interactors, and to the extent that common cells have a larger overall distribution. The co-distribution of constituent species in a community is often used to indicate interaction - whether competitive, trophic, or facilitative (Verberk, 2011; Angelini et al., 2011).

Specific structural aspects of organs with strong 3-D structures, such as the lung, kidney, and other organs that undergo branching morphogenesis, have been modeled with some success (Hartmann and Miura, 2007; Miura, 2008; Oates et al., 2012). However, the cellular organization of other types of organs can be particularly difficult to assess quantitatively. For example, endodermal glandular organ structure is often difficult to perceive, with organizational characteristics that are visible only with the use of molecular or

cellular markers; thus detecting differences in organ structure depends in part on having distinct markers for known functional cellular subsets. Examples of this type of organ include the liver, pancreas, and thymus, all of which are of high clinical importance. Developing a quantitative approach to assess the structure and function of these organs is of direct biomedical relevance.

Here, the thymus was chosen because it is an excellent example of cellular level organization, with a strong connection between organization and function. The thymus consists of developing thymocytes supported by a complex cellular environment containing a variety of resident cell types, including thymic epithelial cells (TECs), dendritic cells, blood vessels, and mesenchymal cells. These cell types comprise multiple microenvironments that direct and support thymocytes to develop from immature progenitors into mature T cells that are both self-tolerant and self-restricted. T cell development in the thymus requires interactions with the thymic microenvironments that provide signals for their survival, proliferation, and differentiation (reviewed in Manley, 2012). Despite their critical role in the generation of cellular immunity and the clinical importance of thymic regeneration, the composition and organization of thymic microenvironments and the mechanisms that promote their proper development and function are not fully understood, in part due to a lack of technical approaches for quantifying tissue-level properties. Therefore, the thymus has many characteristics that make it an excellent system for developing and testing quantitative methods: a diverse cellular composition that can be identified with cell type-specific markers, regional organization that is required for maximal organ function, genetic models with diverse effects on organ composition and function, assays for experimentally inducing organ degeneration and regeneration, and high biomedical relevance. To date, there are no

statistical models of thymic organ structure and function and no established methods for generating one. We seek to overcome this barrier by developing a model of the distribution of cell types that can be used to better understand normal thymic organization and function as well as to evaluate states of disequilibrium.

As an extension of this analysis, we applied our methods to a characterized mouse mutant. We selected a mouse model that contains null mutations in the Aire gene (Anderson et al., 2002) for several reasons. First, the Aire knockout is a well characterized mutant that affects an individual cell population. Null mutations in the Aire gene, which marks a specific subpopulation of mTECs, cause defective negative selection and multi-organ autoimmunity. Second, although Aire mutants have clear defects in stromal function, more subtle defects in stromal composition and organization have been difficult to quantify. Analyses of thymic phenotypes have been performed, and there are differing reports to the effect that Aire mutants have in thymic composition and organization. There is debate as to whether mutant mice mTEC numbers decrease or increase, particularly with respect to K14 (Gillard et al., 2007; Anderson et al., 2002). Other reports have implicated Aire as a regulator of negative selection and regulatory T cell development which has been shown by displaced dendritic cells and reduced expression of T reg cell markers in the Aire mutant thymus (Yano et al., 2008; Lei et al., 2011). These results have functional consequences for different models of Aire function as either a regulator of tissue-restricted antigens in the thymus or as a key modulator of mTEC differentiation and mechanistic implications for autoimmunity in these mice. All of these reports point to the need for more robust quantitative methods to evaluate cellular organization. Although we know Aire is functionally important based upon results from previously published data, we are interested in determining the degree to which Aire

plays a role in maintaining proper organization of the thymus. Through the use of community ecology models, we can analyze what happens after a species, or cell type in this instance, is removed in order to draw conclusions about species interdependency and interactions (Wooten, 2005).

Materials and Methods

Cell Types Included in Analysis

Table 1. Cell types included in initial analysis (7 antibodies for frozen, 7 for paraffin).

Frozen Antibody Set 1	Species	Prep	Cell Type	2ndary
Claudin 3	rabbit	FROZEN	mTEC progenitors	405
CD11c	biotin/hamster		Dendritic	488
PDGFRb	goat		Blood vessels	555
CD25	rat		T cells	647
Frozen Antibody Set 2	**Species**	**Prep**	**Cell Type**	**2ndary**
K14	rabbit	FROZEN	K14+ mTECs	405
Claudin 5	mouse		Blood vessels and some mTECs	488
PDGFRb	goat		Blood vessels	555
Foxp3	biotin/rat		Treg cells	647
Paraffin Antibody Set 1	**Species**	**Prep**	**Cell Type**	**2ndary**
UEA1	biotin/lectin	PARAFFIN	UEA1+ mTEC	405
AIRE	rabbit		Aire+ mTEC	488
p63	mouse		TEC (early stage)	555
Foxn1	goat		TEC(late stage)	647
Paraffin Antibody Set 2	**Species**	**Prep**	**Cell Type**	**2ndary**
UEA1	biotin/lectin	PARAFFIN	UEA1+ mTEC	405
Claudin 5	mouse		Blood vessels and some mTECs	488
Claudin 4	rabbit		mTEC progenitors	555
Foxn1	goat		TEC(late stage)	647

Table 2. Cell types included in final analysis (11 cell types for frozen). We focused on the dataset containing 11 cell types for the remainder of the analysis.

Frozen Antibody Set 1	Species	Prep	Cell Type	2ndary
UEA1	biotin	FROZEN	UEA1+ mTEC	405
CD31	hamster		Blood vessels	488
Claudin3,4	rabbit		mTEC progenitors	555
CD25	rat		T cells	647
Frozen Antibody Set 2	**Species**	**Prep**	**Cell Type**	**2ndary**
K5	rabbit	FROZEN	K5+ mTEC	405
Claudin5	mouse		Blood vessels and some mTECs	488
K14	goat		K14+ mTEC	555
Foxp3	biotin/rat		Treg cells	647
Frozen Antibody Set 3	**Species**	**Prep**	**Cell Type**	**2ndary**
CD11c	biotin/hamster	FROZEN	Dendritic	405
PDGFRb	goat		Blood vessels	488
CD205	rat		cTECs and DC subsets	555

Table 3. Terminology

sampled unit within domain; cell types in quadrats used to find regions of similar composition	spatial	quadrat
linear array of sampled units (quadrats), may be comprehensively or randomly sampled	spatial	transect
physical tissue sample used to assay single-cell thickness composition; 10 microns thick	tissue	section
3 actual/real tissue sections collapsed into 1 section		virtual section
3 virtual sections collapsed into 1 unit		virtual slice

System and Data Collection

Mice

C57BL/6 male mice (N=4) at 5-6 weeks of age were purchased from Jackson Laboratories (Bar Harbor, Maine). Aire deficient mice maintained on a C57BL/6 background (N=2) and normal C57BL/6 male mice (N=2) at 5-6 weeks of age were obtained from Dr. Mark Anderson's lab at the University of California, San Francisco.

Tissue Preparation

Intact thymuses from the Anderson lab were removed and the right and left lobes were separated, placed in OCT compound, and immediately frozen prior to being shipped to the Manley lab. Mice thymuses from Jackson Laboratories were dissected and frozen or embedded in paraffin in the Manley lab.

Antibodies

Primary antibodies used in this work include the biotinylated hamster CD11c (Cat# 117303, BioLegend, 1:150), rabbit Claudin 3 (Cat# 34-1700, Life Technologies , 1:150), rabbit Claudin 4 (Cat# 36-4800, Life Technologies, 1:150), supernatant hamster CD31 (1:50), rat CD25 (Cat# 557425, BD Pharmingen, 1:50), biotinylated UEA-1 (Cat# B-1065, Vector Labs, 1:150), conjugated mouse Claudin5 (Cat# 352588, Invitrogen, 1:150), rabbit Keratin 5 (Cat# PRB-160P-100, Convance, 1:150), biotin rat Foxp3 (Cat# 13577382, eBioscience, 1:50), goat cytokeratin 14 (Cat# sc-17104, Santa Cruz, 1:200; sc-17104, Santa Cruz, 1:800), goat PDGFRb (Cat# AF1042, R&D Systems, 1:150), goat Foxn1 (Cat# sc-23566, Santa Cruz, 1:200), mouse P63 (Cat# sc-8431, Santa Cruz, 1:200), rabbit Aire (Cat# sc-33189, Santa Cruz, 1:50), and rat CD205 (Cat# NLDC-145,Biolegend, 1:150).

Frozen Sample Preparation

The entire left lobe was cut into 10-micron serial sagittal sections using a Leica CM3050 S cryostat. Sections were collected on glass slides and assigned a number corresponding to their location. The middle third of these slides were selected for IHC. Of these middle third, 9 sections from each sample that passed quality control standards were used for IHC. These

sections were fixed in -20 degree C acetone for 20 seconds immediately prior to application of blocking solution (10% donkey serum/PBS) for 30 minutes at room temperature.

Paraffin Sample Preparation

Samples selected for paraffin embedding were fixed with 4% PFA/PBS for 1-2 hours at 4 degrees, washed with 5% sucrose/PBS for 1 hour at room temperature, and dehydrated 1x 30 minutes at 50%, 70%, 90%, 96%, and 100% EtOH. Samples were then dehydrated overnight in 100%EtOH. The samples were placed in Xylenes 2x for 5 minutes and placed in wax 3x for 30 minutes at 60 degrees. The left lobes were cut into 8-micron serial sagittal sections and collected on glass slides. The middle third of these slides were kept for IHC. Immediately prior to staining, samples were washed in Xylenes 2x for 5 minutes, rehydrated for 2 minutes at 2x100%, 95%, and 75% MeOH, and heated in an antigen unmasking solution (10mM Sodium Citrate, 0.05% Tween20) at 95 degrees for 30-45 minutes. Samples cooled down gradually for 20 minutes, and were rinsed in PBS prior to application of blocking solution (10% donkey serum/PBS) overnight at 4 degrees C.

Antibody Staining

Primary antibodies were mixed in PBS and slides were covered with the antibody solution and incubated overnight at 4degrees C. The slides were rinsed with PBS 2x5 minutes and the secondary antibodies were mixed in PBS (1:800) and applied to the slides for 30 minutes. Slides were rinsed with PBS 2x5 minutes, mounted in FluorGel (EMS) and coverslipped. Marker combinations were determined by availability, biological and technical

considerations such as anticipated abundance, distribution of cell types, and reagent compatibility (antibody species of origin).

Microscopy, PTGui, and CellProfiler

Multiple sections from each thymus sample were used in the analysis (9 sections from each sample from the Anderson Lab; 8 sections from WT2, 6 from WT1). The sections were assessed by specific quality control criteria such as lack of section flaws, clarity of image collection, and signal to noise ratio. Sections that passed the QC measures were imaged at 20x as tiled, overlapping quadrats on the Zeiss Axioplan microscope and reconstructed using PTGui. The image produced was optimized by increasing brightness and decreasing background in Adobe Photoshop to reduce background and increase signal intensity so that cell counts could be taken as quickly and accurately as possible in CellProfiler. As the reconstructed image from PTGui was too large to work in CellProfiler, the size of the image was decreased by taking the image width and dividing by 4. CellProfiler was chosen because it has the capability to quickly and accurately identify cells and output locations (x and y coordinates) in a tab-delimited format. Results from the cell counts were manually checked for accuracy and input into R (R Core Team, 2014) for the spatial analysis. Cell counts from serial sagittal sections stained with different marker combinations were collapsed into a single virtual section.

Development of Virtual Sections

Due to technical limitations we can currently label a tissue section with 3-4 markers; to maximize the amount of information available for analysis, multiple serial sections were

stained with different sets of markers and collapsed into a single virtual section (Figure S1). Virtual sections were collapsed to create a virtual slice.

Identification of Scale

As thymic epithelial cells vary in size from X to X', we considered the problem of how to generate sampled regions (quadrats) that contained sufficient information about the local community of cells so that these regions could be classified (Figure S2). In other words, we need to have enough data in each quadrat to have statistical power for clustering and we need enough quadrats to improve the spatial resolution. We accomplished this by finding the *relevé*, the quadrat size that encompasses the minimal area yet contains the majority of diversity (Barbour, 1987) . We produced a rarefaction curve (Figure S3) for the different quadrat sizes and found that quadrat sizes from 0.0005 to 0.0025 mm^2 would provide the best spatial resolution and statistical power. One of the benefits of our analysis is that our sampling technique covers the entire span of a thymus section; the use of antibody staining identifies all of the cell markers included in the analysis that cover the tissue section. Ecologists, in contrast, must rely upon measures which randomly sample their region of interest; covering an entire community and identifying all species would take too much time and too many resources. Our method does not suffer from these limitations; we are able to essentially take a snapshot of our entire 2D "community" and use it to identify and study cellular patterns at a fixed point in time. The sampling technique most similar to ours is remote sensing, which is able to provide a synoptic view and deliver information over large areas at high levels of detail (Nagendra, 2001; Immitzer et al., 2012).

Spatial Analysis with R

Files containing the X and Y coordinates from CellProfiler were uploaded into R and reformatted (package gplots, Warnes et al., 2015) into a community matrix. The K-means function in R's stats package was used to cluster the data according to cell type, abundance, and location. As no likelihood model exists for the interaction and spatial distribution of cell types, we bounded our evaluation of K (the number of clusters identified from data) at a minimum of 3 (putatively identifying cortical, medullary, and "other" spatial regions, the "other" generally including the space outside of the thymic section) and a maximum of 10. As K increases, the general effect is continued subsetting of regions identified at K-1; thus our criterion for consideration of a number of K regions depended on a measure of differential composition of the regions. The Bray-Curtis function (package Vegan, Oksanen et al., 2015) was used to calculate the Bray-Curtis dissimilarity index (Bray and Curtis, 1957) for the different regions produced from K-means clustering. When an increase in K led to spatial regions separated by Bray-Curtis values < 0.1, we considered the additional region of little biological significance.

To evaluate the spatial arrangement of the K regions identified, we recognized that each region is often comprised of multiple areas that share similar cellular composition. A 'connected component labeling' method was used to identify individual areas of type K that are spatially separated. We used the ConnComp function (package SDMTools, VanDerWal et al., 2014) package to calculate and produce the connected components. This approach then enables the spatial parameters of each component (area, centroid, etc.) to be evaluated with respect to other components of the same type as well as their distribution relative to components of other types.

Results

K-means clustering reproduces previously identified regions and identifies cryptic information

We used the K-means clustering algorithm to identify geographical clusters of cell types based on compositional similarities and abundance. We used a range of K values to compare to previously identified regions of the thymus and to identify cryptic organization. For example, K=3 reproduced the cortical and medullary regions (as well as the blank space outside the thymus, which we chose to keep in our analysis instead of removing because it provided a way for us to validate our results) which was expected since a majority of the antibody markers correspond to cell types specifically located within these regions. We increased K to see if we could reproduce the CMJ or subcapsular region since we do not have any antibodies that specifically mark these areas. With K=4, we see formation of the outside of the section, the cortex, and the subdivision of the medulla (which we called medulla 1 (M1) and medulla 2 (M2)), which could either represent the CMJ or a new unidentified functional region. Increasing K further results in clusters forming throughout the medulla and eventually throughout the cortex. At higher K values, we do eventually see formation of an apparent subcapsular region. We focused on K=4 for the remainder of the study because we were interested in the formation of the 2 clusters (subdivision) in the medullary region (Figure 1).

We repeated these steps with another wild-type mouse and 2 mutant mice (lacking Aire, a specific medullary thymic epithelial cell subset) and saw different cluster patterns in the M1 and M2 regions (Figure 2); this strongly shows we need to increase sample size in order to determine which patterns are real (of course, it is possible that all of these patterns are real

and that there is a fair amount of variability between individual thymuses). We decided to focus on our first wild-type sample and proceed with the development of methods (we picked the first wild-type sample since the cluster pattern was similar to that seen in previous datasets that only contained 7 cell types).

Our next step was to determine by what factors (cell types and abundances) the K-means clustering algorithm was grouping quadrats into clusters. We produced graphs showing the abundance of cell types within each cluster (Figure 3). We were also able to produce graphs showing the proportion of cell types across clusters (Figure 4). Since our dataset contained a larger number of markers for the medullary region, we produced density plots to make sure the abundance of cell types was not the sole driving factor of the clustering (Figure S4). Although the amount of cells does influence the clustering, the results were not completely dependent on density. We also normalized our data to completely remove the effect of abundance. However, we mainly saw subdivision of clusters in the cortex (Figure S5). Since our dataset is rich in medullary markers, we choose to focus on clustering in the medulla, which is not possible with the normalized data. To evaluate the effect of individual cell markers, we removed specific cell markers, in this instance K14 (pan-medullary marker), from the dataset to see how clustering was affected (Figure 5). By removing specific markers from our dataset, we are able to manipulate our dataset and focus on the cluster formation of certain cell types that were previously undetectable when the more dominant cell types were present.

Bray-Curtis Identifies Degree of Dissimilarity

We used the Bray-Curtis dissimilarity index to determine the degree of compositional dissimilarity between clusters within a thymus section (K=4), clusters in different sections in the same individual thymus (N=3), and clusters between sections in different thymuses (N=2). The Bray-Curtis index is bounded at 0 and 1, representing complete similarity in composition to complete lack of similarity; these values can be grouped in a dendrogram using a neighbor-joining algorithm for easier visualization. The K clusters produced analytically tended to group together by recognized "geographical" region (cortex with cortex, etc), with the exception of M1 for our first wild-type sample, which grouped more closely with M2 (Figure 6). We will need more samples to determine what this means, but the general trend from these results indicates that the clusters produced from K-means are compositionally very similar to other wild-type sample clusters. We repeated these steps with mutant mice and saw a similar result: the mutant and wild-type clusters both grouped with regions of geographical similarity (Figure 7). **This result indicates that, compositionally, Aire mutant and wild-type thymuses are not significantly distinct.**

Connected Component Labeling

Next, we wanted to determine if mutant and wild-type samples differ spatially with respect to the location of the M1 and M2 clusters. The results from the K- means clustering show that for one of our wild-type mice, M2 is inset in M1, while for one of our mutant mice, M1 and M2 appear to be more adjacent (Figure 8). Based upon this pattern, we would expect the average distance between clusters to be greater for the mutant thymus than for the wild-type; the centroids for the wild-type thymus should be overlapping or very close together since

the red cluster is surrounding the yellow cluster while the centroids for the mutant should be further apart since the red and yellow clusters are adjacent to each other. We used connected component labeling to identify individual cluster components and calculated the centroid of these components (Dillencourt et al., 1992). The connected component labeling algorithm correctly identified cluster components for the mutant sample (Figure 9, asterisk indicates centroid) but was imperfect in its ability to detect individual cluster components due to technical constraints for the wild-type sample (Figure 10, the red clusters were not completely separate from other red clusters in the wild-type) which resulted in the skewed placement of the centroid. In order to overcome this misidentification of cluster components, we manually selected individual cluster components, ignoring those that were identified incorrectly. Cluster components for the wild-type thymus were between 2-3 pixels apart (100-150 microns), while the distance for the mutant thymus cluster components varied between 5-6 pixels (250-300 microns). We repeated our experiment with a second mutant and wild-type, and found that the organization of the clusters was considerably different compared to our initial samples. Given our low sample size, constrained by the development of methods and funding, we do not yet have the statistical power to determine if the pattern of organization in the first mutant and wild-type samples are consistent among individual thymuses, or if there is a certain degree of variability that can be expected and quantified.

Discussion and Conclusion

Our approach is designed to take advantage of the cellular complexity of the thymus in order to provide a sensitive statistical method to measure organizational changes in organ state. We used immunostaining on sagittal serial sections of a wild-type mouse thymus to identify distinct cellular subsets, followed by the use of novel computational approaches to

quantitatively identify known and cryptic cellular spatial relationships. These methods are designed to replace the largely descriptive histological and IHC-based analyses currently available in the field, thus providing a more accurate and unbiased view of thymus biogeography. Our experimental approach is essentially 2-D because it involves the use of antibody staining and detection on individual, single-cell layer thick 2-D thymic sections. Although the 3-D nature of the stromal network is an important aspect of its functionality (van Ewijk et al., 1999) , a 2-D picture will provide an informative, if not complete view of the thymus. After all, the qualitative assays currently used are all viewed in 2D sections and are the basis of our current views of thymus organization and function.

The first objective of our analysis involves the development of a baseline model of a healthy, 1-month old thymus, assuming that organizational and compositional patterns are similar between individuals. We must first understand how general observed patterns compare across individuals of the same age and genotype before we can make comparisons with abnormal thymuses. K-means clustering provided the first step in making these comparisons. K- means clustering allowed us to test the structure of the thymus in a statistical, as opposed to descriptive, framework that provided support for the number of compositionally distinct thymic regions. Unlike descriptive approaches, K-means clustering is automatic and dependent on objective criteria for sorting the functional cell types in a spatially explicit way. First, we used the K-means clustering algorithm to identify geographical clusters of cell types based on compositional similarities, position, and abundance. This clustering algorithm generates a posterior probability that any subsample (quadrat) belongs to each of the different K groups, while maximizing the ratio of compositional differences between groups to compositional variation within groups. This

approach has previously been successful in delineating distinct cell sub-populations in tissue samples (Veronika et al., 2009; Di Cataldo et al., 2009). We compared the results to previously identified regions (which were identified through descriptive approaches) of the thymus to see if we recreated the same regions or discovered new, functionally distinct regions or subregions. Our results supported previous discussion of the geography of the thymus. For K=3, we reproduced the cortex, medulla, and the empty blank space surrounding the thymus. We then increased K to see if we could identify potentially cryptic organization, such as functional microenvironments.. For K=4, we saw subdivision in the medulla which could either be the CMJ or a new, previously undescribed subregion that could have functional significance. By further increasing K, we began to see further subdivision of the medulla, as well as clustering throughout the cortex. To an extent, as K goes up it may be that spatial weighting is important, but we choose to avoid the problems that weighting brings up; it may also be possible that some of the areas defined as K approaches 10+ are legitimate microenvironments and this will be addressed in a later paper. We focused on K=4 for the rest of the analysis in order to further develop our methods. We produced graphs which showed the number of each cell type within each cluster as well as graphs which allowed for comparisons of cells between clusters. Although these visuals were useful for understanding how the K-means clustering was being performed, the goal of this project was to move away from descriptive analyses and more towards statistical methods.

In ecology, the identity and relative abundance of species can be compared across communities using various dissimilarity indices (Magurran, 2004). The Bray-Curtis dissimilarity index in particular is a well characterized statistical method used to quantify the compositional dissimilarity between two sites, based on counts at each site (Bray and Curtis,

1957). We applied the Bray-Curtis index to our samples to determine the degree of dissimilarity between clusters produced from K-means for different regions in a section, between sections in the same individual, and between samples in different individuals. We organized the results into a dendrogram for easier visualization and found that the clusters which corresponded to similar geographical regions grouped together (for example, the cortex from one sample grouped with the cortex from another sample), which indicated that the wild-type thymuses are compositionally similar. The Bray-Curtis values produced ranged from 0.2-0.9, which is similar to the ranges seen between comparisons of salt marshes (Carlisle, 2004), indicating that the level of diversity within a small, 7 (h)x2 (w) mm section is strikingly similar to that of a coastal ecosystem.

Our goal was not only to identify whether groups of 2 or more cell types have similar relative distributions (between regions or within the same regions in different sections/samples), but also to quantify the extent to which mutants or diseased thymuses are significantly less structured than wild-type. We applied the same approaches described above to a mutant mouse thymus in order to determine if there are any detectable organizational differences compared to the wild-type. The parallel in ecology is removal of a species from a community, which can lead to two distinct outcomes: the relative species abundance may not be affected if species are independently utilizing available resources, or the relative abundance of species may change dramatically, suggesting strong interactions depending on missing components. This type of species is known as a keystone species or foundational species, and its availability has strong implications on the proper functioning and survival of the community it belongs to. The results from both the K-means clustering algorithm and the

Bray Curtis dissimilarity index failed to detect any significant compositional differences between the mutant and wild-type samples.

Next, we checked for distributional differences with respect to cluster position since the location of M1 relative to M2 appeared different between the mutant and wild-type samples. We used connected component labeling to group connected components within an individual cluster together. We were then able to calculate the centroid of each cluster component and calculate the shortest distance between cluster components within different clusters. For the wild-type mouse thymus, we expected the M1 and M2 centroids to overlap, with a small distance between the centroids since M1 appears to surround M2. For the mutant sample, we expected to see a larger distance between the centroids since the M1 and M2 clusters appeared to be more adjacent to each other. CCL was unable to detect individual cluster components due to technical constraints for the wild-type sample which resulted in the skewed placement of the centroid. In order to overcome this misidentification of cluster components, we manually selected individual cluster components, ignoring those that were identified incorrectly. The distance between clusters in the wild-type sample was on average between 2-3 pixels (100-150 microns), while the distance between the mutant clusters was closer to 5-6 pixels (250-300 microns). We repeated our experiment with a second Aire mutant and wild-type thymus and found the organization of clusters to be variable between all 4 samples. Therefore, it is essential that we increase sample size before we draw conclusions about organization between wild-type and mutant thymuses. Our sample size is low due to method constraints and limited funding and because of this, we do not yet have the statistical power to determine if these patterns are real. Future work will seek to remedy this.

Future Directions

There are many different directions we could take this project. For instance, we could examine how removal of specific cell types from the dataset will alter clustering as well as the statistical analyses. We could further manipulate K and examine in further detail the clusters that appear when K is increased. It is possible that these clusters may have functional significance (for instance, they could be microenvironments that have not previously been identified). We are also interested in characterizing the organization of other thymic mutants and thymuses that are undergoing involution. This technique could be used to evaluate clinical interventions for thymic regeneration. The methods developed in this study could easily be further modified or developed to include other ecological or statistical methods, to the point that they could also be used to quantify the organization of other organs as well. For instance, the use of the 2-point correlation function (commonly used in astronomy to describe the distribution of galaxies) could be used to provide information on both the distribution of a cellular population within the thymus as well as its relative distribution with respect to other cell types.

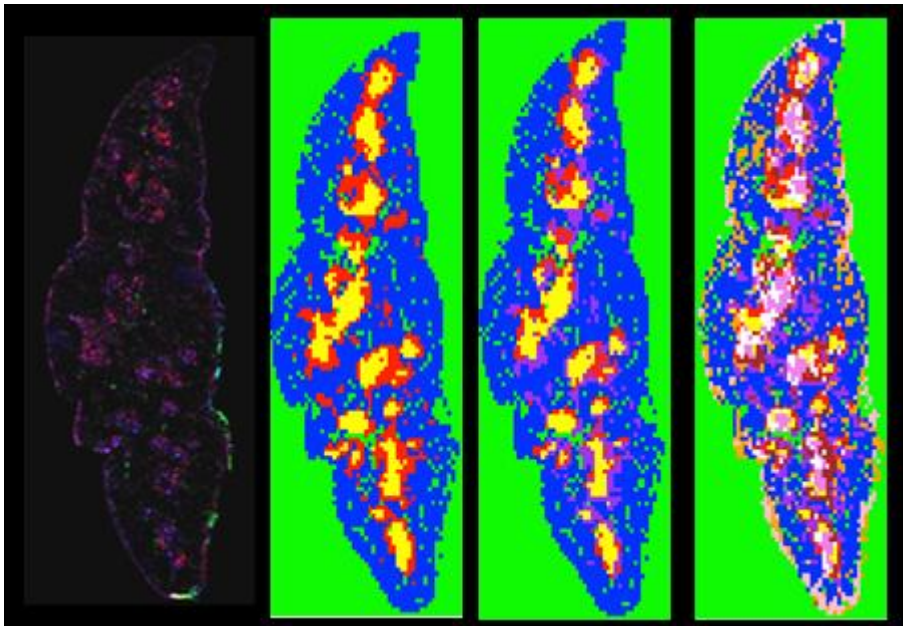

Figure 1. Results of clustering using K-means. Left: Recreated physical tissue sample stained with 4 cell markers (CD31, UEA1, Claudin3,4, CD25). Right: With K=4, we saw clustering in the outside of the section (green), the cortex (blue), and the medulla (yellow and red). Increasing K further (K=5, K=10) produced clusters in the cortex and medulla.

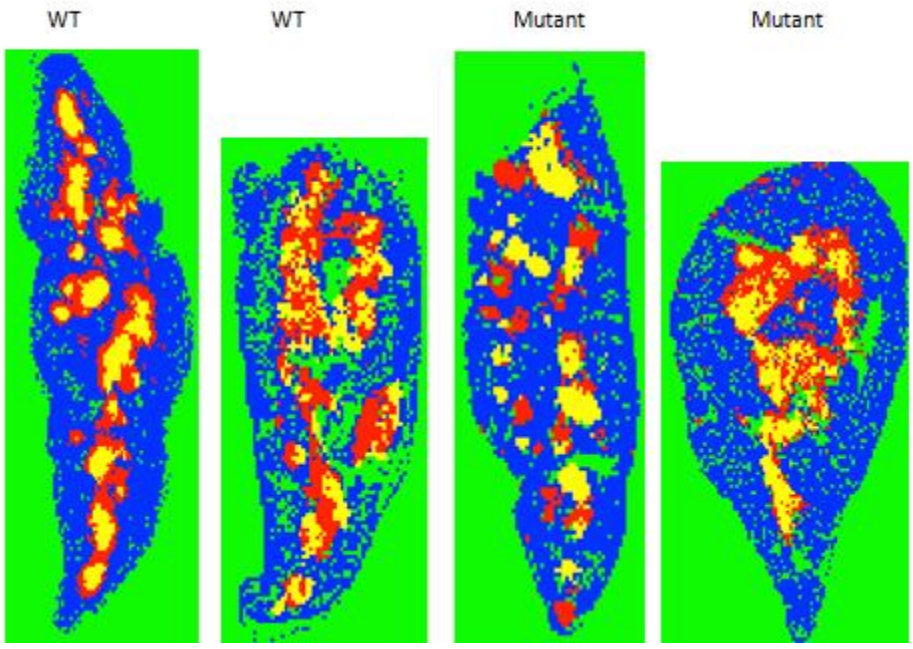

Figure 2. K-means clustering for 2 wild-type and 2 mutant samples (K=4).

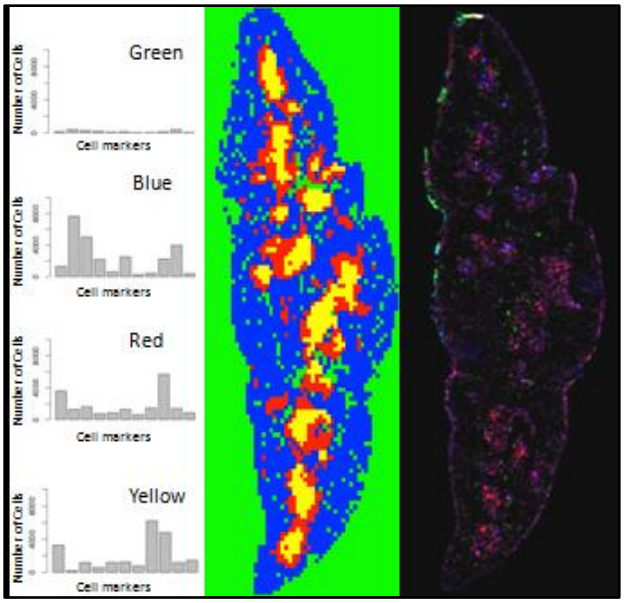

Figure 3. Abundance of 11 cell types within clusters. The blue cluster has an abundance of several cell types, all of which correspond to the cortex. The red and yellow clusters have similar patterns of abundance with subtle differences; the main difference is there is an abundance of one cell type (K14) in the yellow cluster that isn't as abundant in the red cluster.

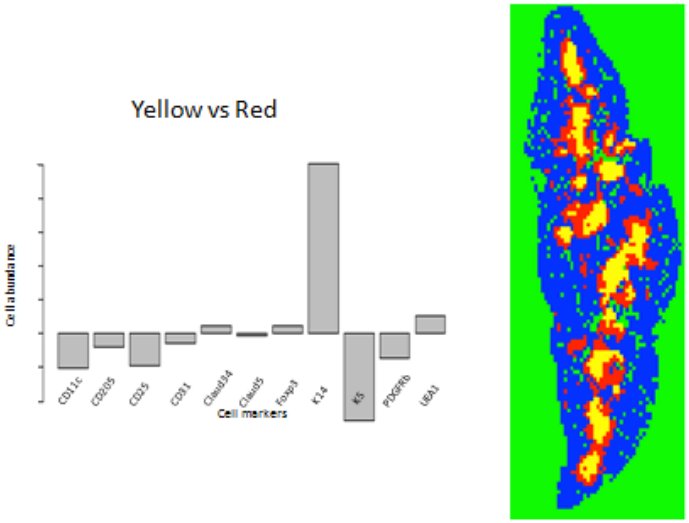

Figure 4. Comparison of cell types between clusters. We focused on the proportion of cell types between the two medullary regions (M1 and M2), represented by yellow (top portion of graph) and red (bottom portion of graph). According to the results, K5 (mTEC subset marker) is the most abundant cell type in the yellow cluster and K14 (mTEC subset marker) is the most abundant in the red cluster.

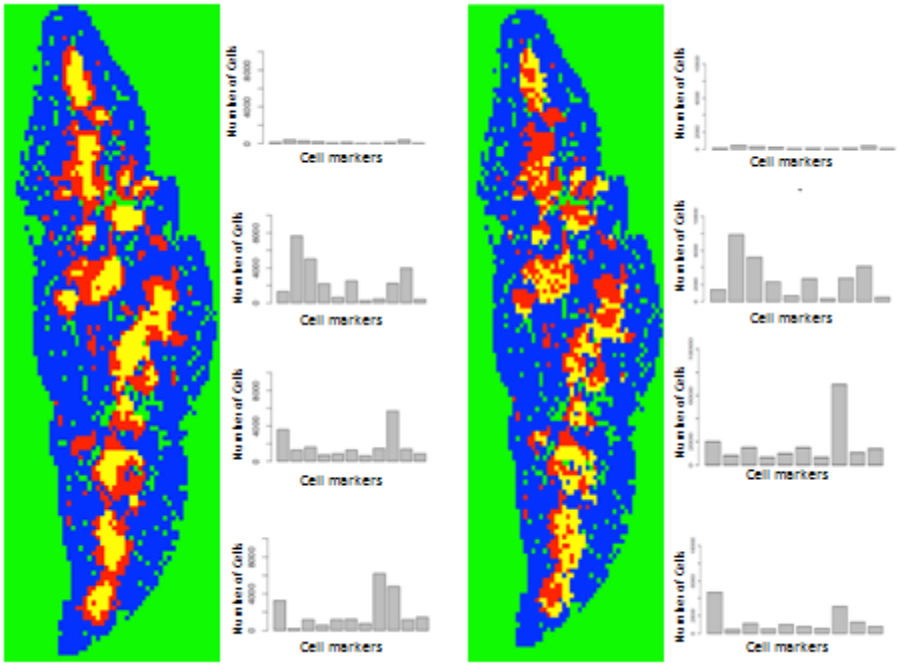

Figure 5. Comparison of 2 virtual sections; left section shows results produced using a dataset with all 11 markers and right section shows results when we removed a marker from the dataset (K14, pan-mTEC marker). Left section shows clustering in the CMJ. Right section shows more interspersed clustering in the medullary region, indicating that the presence of K14 is one of the driving factors of the clustering in the first section.

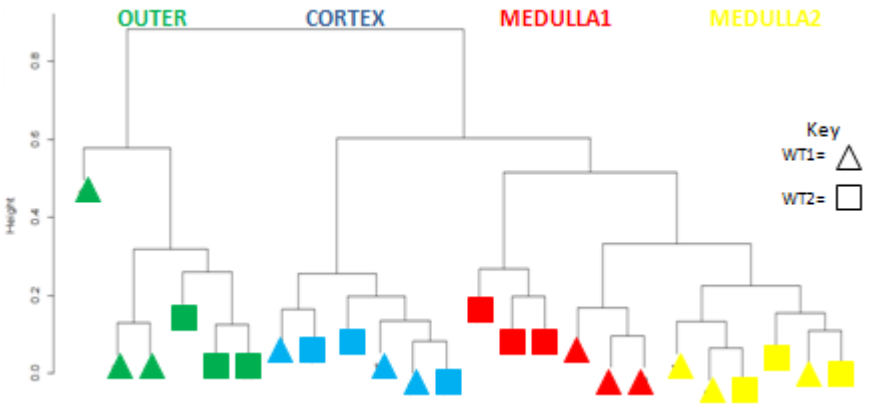

Figure 6. Dendrogram of Bray-Curtis results for wild-type (N=2) thymus samples. The different colors correspond to the different clusters produced from K-means clustering.

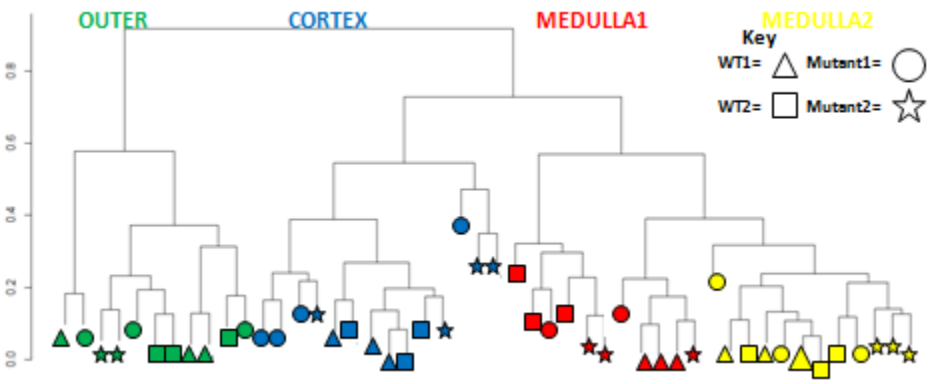

Figure 7. Dendrogram of Bray-Curtis results for wild-type (N=2) and mutant (N=2) thymus samples.

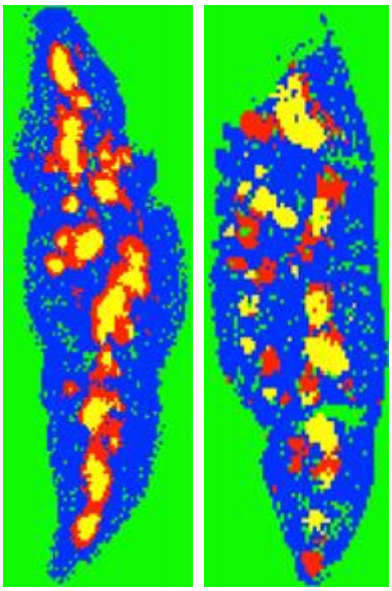

Figure 8. Comparison of clusters produced for wild-type and mutant samples (K=4). The location of location of the red and yellow clusters in the wild-type thymus appears different to that of the mutant thymus.

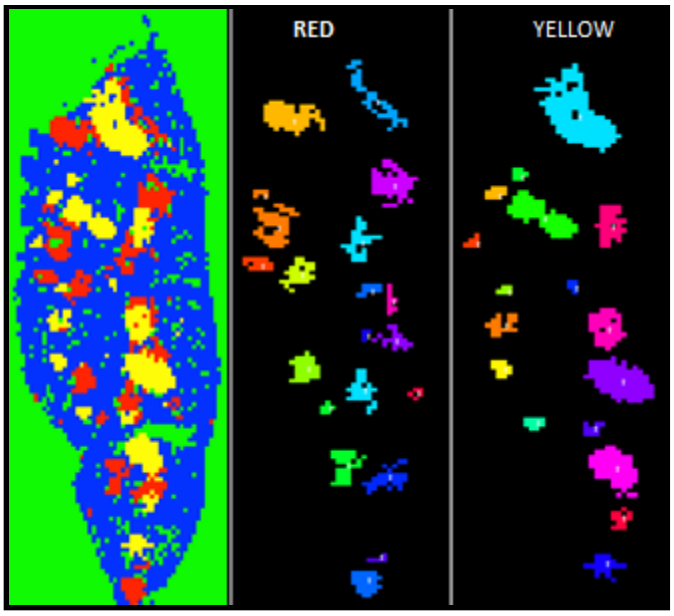

Figure 9. Comparison of cluster components created from connected component labeling (CCL) for our first mutant section. Left: Virtual section produced from K-means clustering (K=4). Middle: Cluster components of M1 (outer medullay region, red clusters in virtual section) represented by different colors. Right: Cluster components of M2 (innter medullary region, yellow clusters in virtual section).

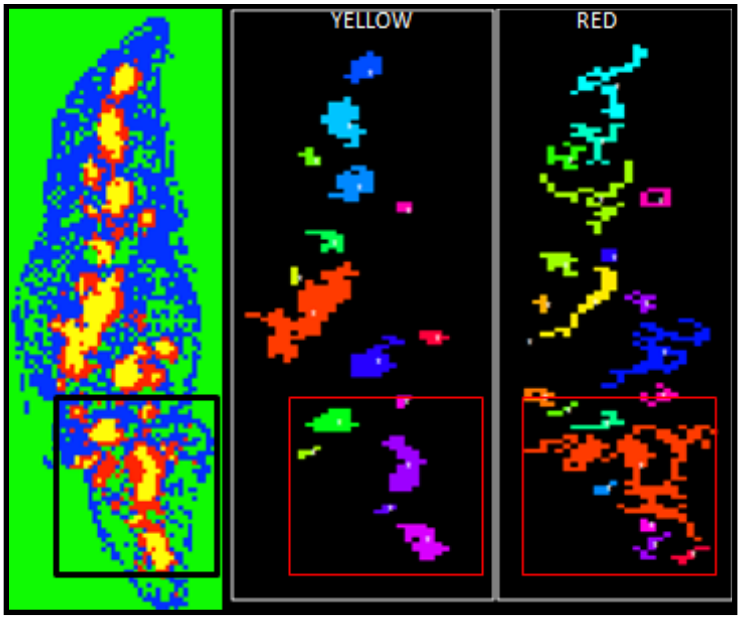

Figure 10. Comparison of cluster components created from CCL for our first wild-type thymus. Left: Virtual section produced from K-means clustering (K=4). Middle: Cluster components of M2. Right: Cluster components of M1. Boxes indicate region where CCL incorrectly identified cluster components for the red cluster.

References

Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP, et al. (2002) Projection of an immunological self shadow within the thymus by the aire protein. Science 298: 1395-1401.

Angelini C, Altieri AH, Silliman BR, Bertness MD(2011) Interactions Among Foundation Species and Their Consequences for Community Organization, Biodiversity, and Conservation. Bioscience 61 (10): 782–89.

Barbour MG, Burk JH, Pitts WD (1987) Terrestrial Plant Ecology. Benjamin/Cummings Publishing Company.

Bray JR, Curtis JT (1957) An Ordination of the Upland Forest Community of Southern Wisconsin. Ecology Monographs.

Carlisle BK, Baker JD, Hicks AL, Smith JP, Wilbur AR (2004) Cape Cod Salt MarshAssessment Project; Final Grant Report, Volume 2: Response of selected salt marsh indicators to tide restriction 2000-2003. Boston, MA. Massachusetts Office of Coastal Zone Management.

Cataldo S, Ficarra E, Macii E (2009) Automated Discrimination of Pathological Regions in Tissue Images: Unsupervised Clustering vs. Supervised SVM Classification. In: Fred A, Filipe J, Gamboa H, editors. Biomedical Engineering Systems and Technologies: Springer Berlin Heidelberg. pp. 344-356.

Dillencourt MB, Samet H, Tamminen M (1992) A General Approach to Connected-Component Labeling for Arbitrary Image Representations. Journal of the ACM 39 (2): 253–80.

Gillard GO, Dooley J, Erickson M, Peltonen L, Farr AG (2007) Aire-dependent alterations in medullary thymic epithelium indicate a role for Aire in thymic epithelial differentiation. J Immunol 178: 3007-3015.

Hartmann D, Miura T (2007) Mathematical analysis of a free-boundary model for lung branching morphogenesis. Mathematical medicine and biology : a journal of the IMA 24: 209-224.

Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton, NJ: Princeton University Press.

Immitzer M, Atzberger C, Koukal T (2012) Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. Remote Sensing 4 (9): 2661-2693.

Krebs CJ (1998) Ecological Methodology, 2nd ed. Menlo Park, CA: Benjamin Cummings.

Lei Y, Ripen AM, Ishimaru N, Ohigashi I, Nagasawa T, et al. (2011) Aire-dependent production of XCL1 mediates medullary accumulation of thymic dendritic cells and

contributes to regulatory T cell development. J Exp Med 208: 383-394.

Magurran AE (2004) Measuring Biological Diversity. Malden, MA: Blackwell Publishing.

Manley NR, Richie ER, Blackburn CC, Condie BG, Sage J (2012) Structure and function of the thymic microenvironment. Front Biosci 17: 2461-2477.

Miura T (2008) Modeling lung branching morphogenesis. Current topics in developmental biology 81: 291-310.

Nagendra H (2001) Using remote sensing to assess biodiversity. International journal of remote sensing. 22(12): 2377-2400.

Oates AC, Morelli LG, Ares S (2012) Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. Development 139: 625-639.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015). vegan: Community Ecology Package. R package version 2.2-1. http://CRAN.R-project.org/package=vegan

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

VanDerWal J, Falconi L, Januchowski S, Shoo L, Storlie C. Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. http://www.rforge.net/SDMTools/l

van Ewijk W, Wang B, Hollander G, Kawamoto H, Spanopoulou E, et al. (1999) Thymic microenvironments, 3-D versus 2-D? Semin Immunol 11: 57-64.

Verberk W (2011) Explaining General Patterns in Species Abundance and Distributions. Nature Education Knowledge 3: 38.

Veronika M, Evans J, Matsudaira P, Welsch R, Rajapakse J (2009) Sub-population analysis based on temporal features of high content images. BMC Bioinformatics 10 Suppl 15: S4.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumleyn T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2015). gplots: Various R Programming Tools for Plotting Data. R package version 2.16.0. http://CRAN.R-project.org/package=gplots

Wootton JT (2005) Field parameterization and experimental test of the neutral theory of biodiversity. Nature 433: 309-311.

Yano M, Kuroda N, Han H, Meguro-Horike M, Nishikawa Y, et al. (2008) Aire controls the differentiation program of thymic epithelial cells in the medulla for the establishment of self-tolerance. J Exp Med 205: 2827-2838.

Appendix

Supplementary Material

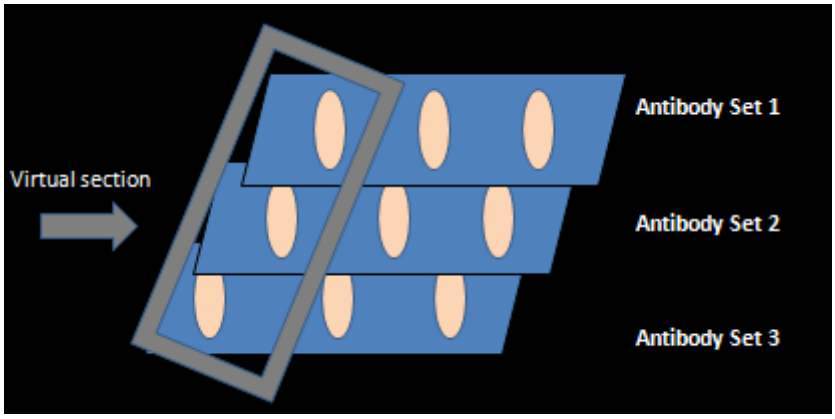

Figure S1. We collapsed 2D sections stained with different antibody sets into a single 2D "virtual" section.

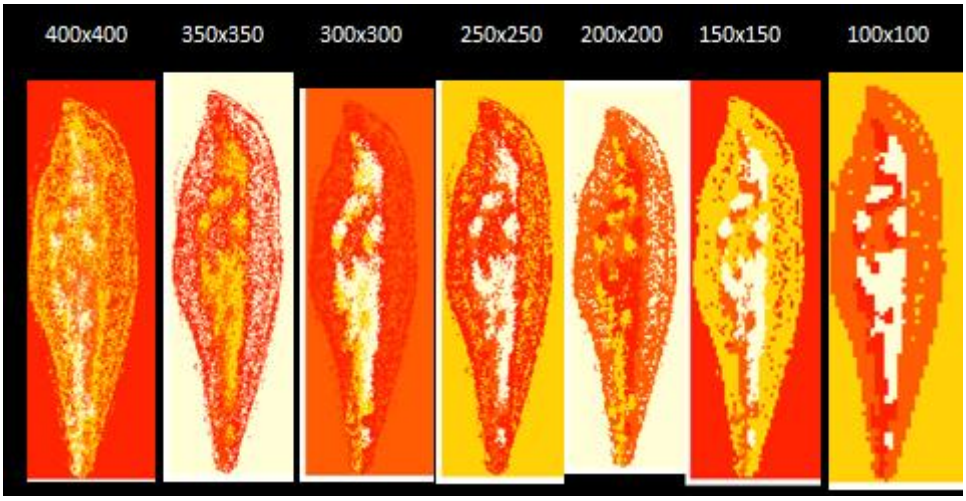

Figure S2. We produced sections using different matrix sizes (listed above the section) to see how the resolution differs between large and small bin sizes (sections produced are from frozen dataset containing 7 cell types).

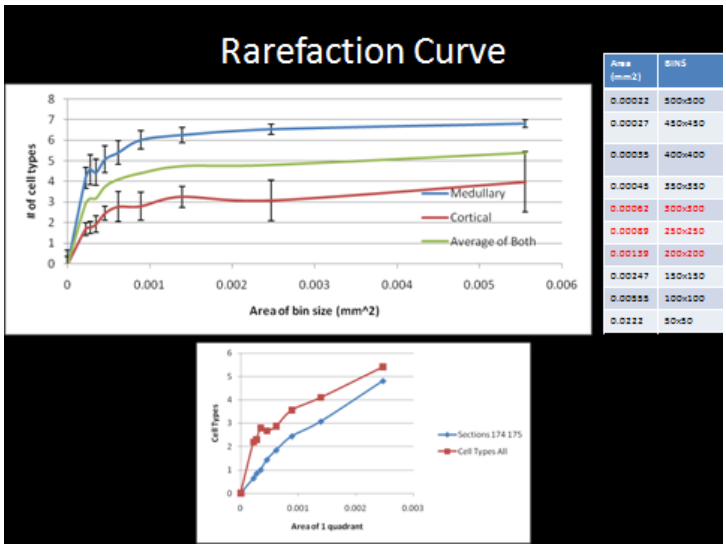

Figure S3. Rarefaction curve (produced from frozen dataset containing 7 cell types) used for determination of optimal quadrat size. The top graph sampled the medulla and cortex separately and took the average of both to produce the 3 different lines. The bottom graph sampled throughout the section randomly (which is less biased and more similar to how ecologists sample communities).

Figure S4. Right: Results of K-means clustering. Left: Density plot of cell counts.

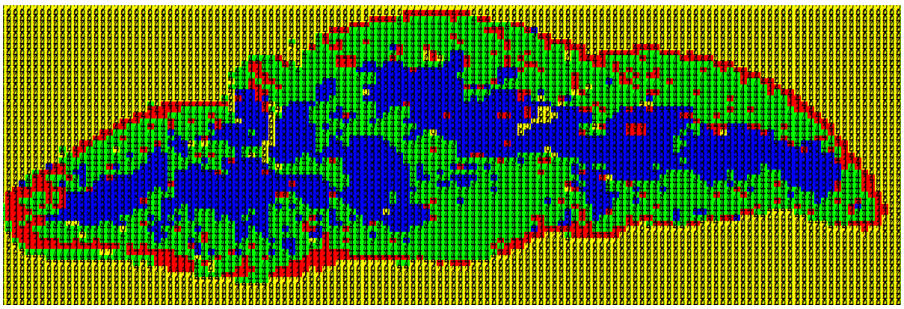

Figure S5. Results of K-means clustering using data that has been normalized. We see reproduction of cortex, medulla, subcapsular region, and the outside of the section.