

Ecology of the Thymus

Mary Browning, John Wares, Nancy Manley

10 December, 2014

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, tissues, and cells. Without quantitative methods for assessing organization, a significant aspect of organ and tissue biology is inaccessible to analysis and contrast. Fortunately, a striking comparison can be drawn between the interactions that occur within organs 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 of other species. The presence or absence of one type may alter the functional role and dominance of another, or the overall service the community provides to the larger system (e.g., ???). By comparing cell types within a biological organ to species in an ecosystem, quantitative methods used for ecological analyses can be directly applied. Through analysis of the exact spatial location and co-location of component cell types, in an approach that uses the analytical toolkit of community ecologists and biogeographers, we gain a better understanding of which interactions can be validated by independent approach and analysis, as well as which interactions or even cell types we are likely to be need more information about.

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). 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 codistribution 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 endemcity or dramatically shifted abundance), the colocalization 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 (this is not tautological, this follows from the first: you cannot be rare and widespread in a functional sense). Our overall goal is development of a cellular ecology, where we can understand interactions at a quantitative level that does not yet exist in developmental biology, particularly in this system. We seek to create a map of a healthy adult organ using cell location and abundance, and then to spatially characterize the organ by applying ecological theory directly to the produced map (something that has never been done before). The results from this analysis can be used to make comparisons with organs in states of disequilibrium, such as diseased or mutant organs, in order to see what organizational differences occur.

We performed this analysis on our model organ, the thymus. The thymus was chosen because it is an intriguing example of cellular level organization, with a strong connection between organization and function. The thymus consists of developing T cells supported by a complex cellular environment containing a variety of resident cell types, including thymic epithelial cells (TECs), dendritic cells, vasculature, 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 [1]. 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 and theoretical approaches for quantifying tissue-level properties. Thus, the thymus has many characteristics that make it an excellent system for developing and testing quantitative modeling: 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 quantitative models of thymus organ structure and function and no established methods for generating one. We seek to address this by developing a quantitative theoretical model of the organization of cell types that can be used to better understand thymic function as well as evaluate diseased states. Our approach uses immunostaining on sagittal serial sections of a wildtype mouse thymus in order to identify distinct cellular subsets, followed by the use of novel computational approaches in order to quantitatively identify known and cryptic cellular spatial relationships. We then take this approach one step further by comparing a wildtype and a mutant (Aire knockout) mouse thymus in an effort to identify differences in organizational characteristics.

Methods

Terminology

I suggest an explanatory box for readers so that quadrat, section, composite, etc. can all be explained clearly in one place.

quadrat	spatial	sampled unit within domain; cell types in quadrats used to find regions of similar composition
transect	spatial	linear array of sampled units (quadrats), may be comprehensively or randomly sampled
layer	tissue	actual tissue sample used to assay single-cell thickness composition

1. Where the mice come from and other necessary information about the Aire mutant line (Mary, Nancy)
2. Information about the antibody stains, the frozen versus paraffin protocols
3. Staining protocol and visualization with Cell Profiler: you need to explain whatever settings/parameters were used to obtain your data files, including how overlapping sections were lined up.
4. Data Transformation
5. Quadrat Selection and Rarefaction: you need to explain how you did this, and the next chunk can include the code you used to achieve this. That way, our results can include one plot and a brief discussion about the selection of that size quadrat and what it means for inclusion of cell types in a given quadrat
6. Now get into the methods you already have in your R script, first setting K=2 and removing all quadrats that fall outside of the thymus from further analysis. You can do this using the subset command I think. We can then talk about more ways to maintain the identity of a given cluster of spatial quadrats (identity being the type of mouse, the section of thymus, and so on)

6b. bray curtis

6c. significant diffs between cell types?

7. also Andrew Sornborger's approach for separating out the geometry/geography of clusters., and/or using connected component labeling to identify size, centroid, distance of centroids among clusters
8. jackknifing : removing one cell type at a time

CellProfiler outputs X&Y coordinates in an excel file (which also contains other unnecessary info). I delete the unnecessary info and add points ((0,0), (0,X), (X,0), (X,X)) to the X and Y columns in order to make a

square so that the bin sizes in the matrix will be a consistent size. I save this file as a .txt. It is now ready to run through my script. The next step is to transform my data from individual .txt files into a community matrix that contains all the cell types. The community matrix is then used for Kmeans clustering.

Read in datafile generated from CellProfiler (1 cell type per file)

```
celltypes<-c("CD11c","CD25","CD31","CD205","Claud5","Claud34","Foxp3","K5","K14","PDGFRb","UEA1")
numcells<-length(celltypes)
layers<-3
numfiles<-(numcells*layers)

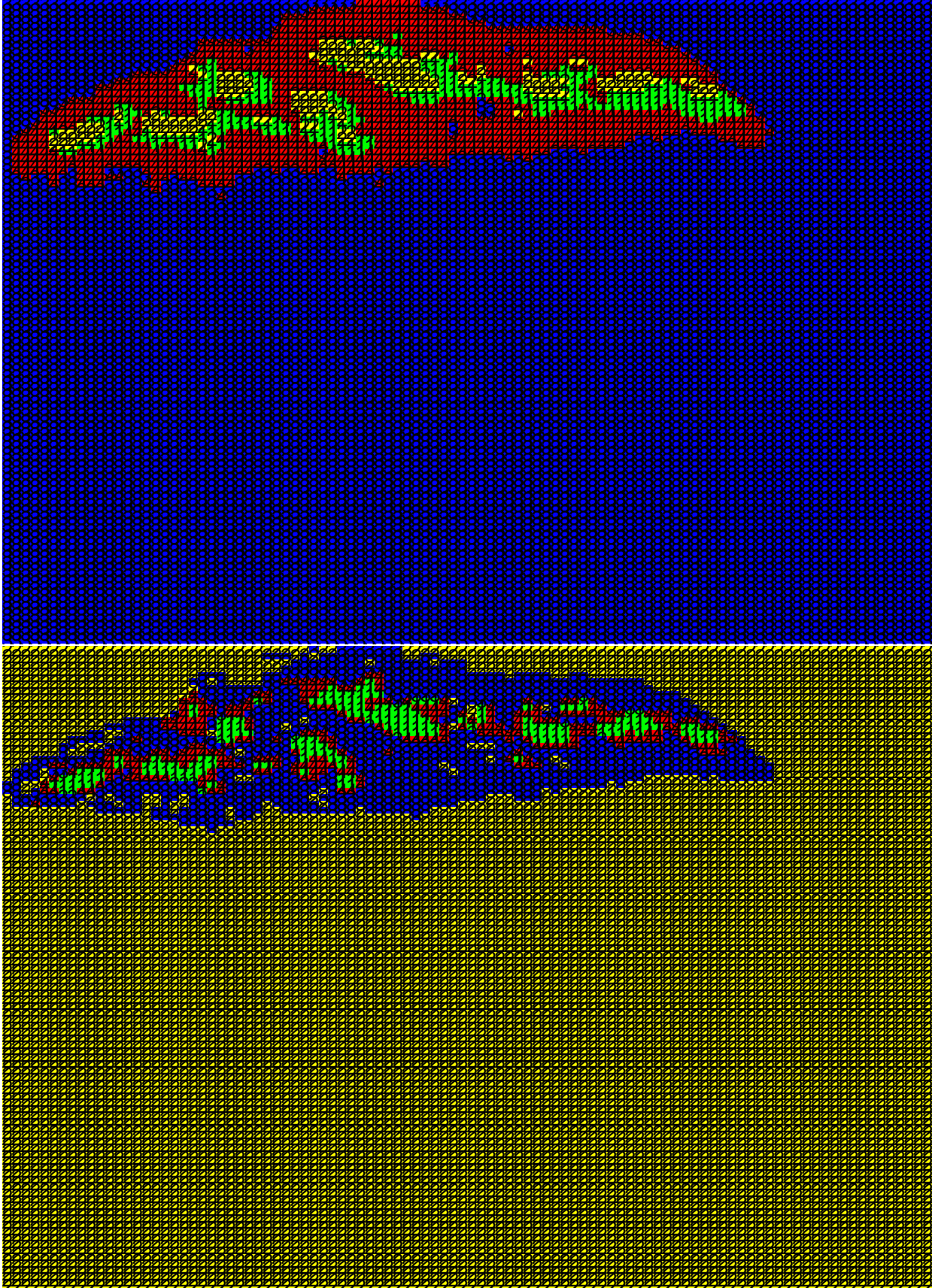
#setwd("~/Desktop/ThymusGit/Datasets/Data with 11 cell types/WT1")
#WT1 <- lapply(Sys.glob("*.txt"), read.table) #this puts ALL from this directory into a list of dataframes

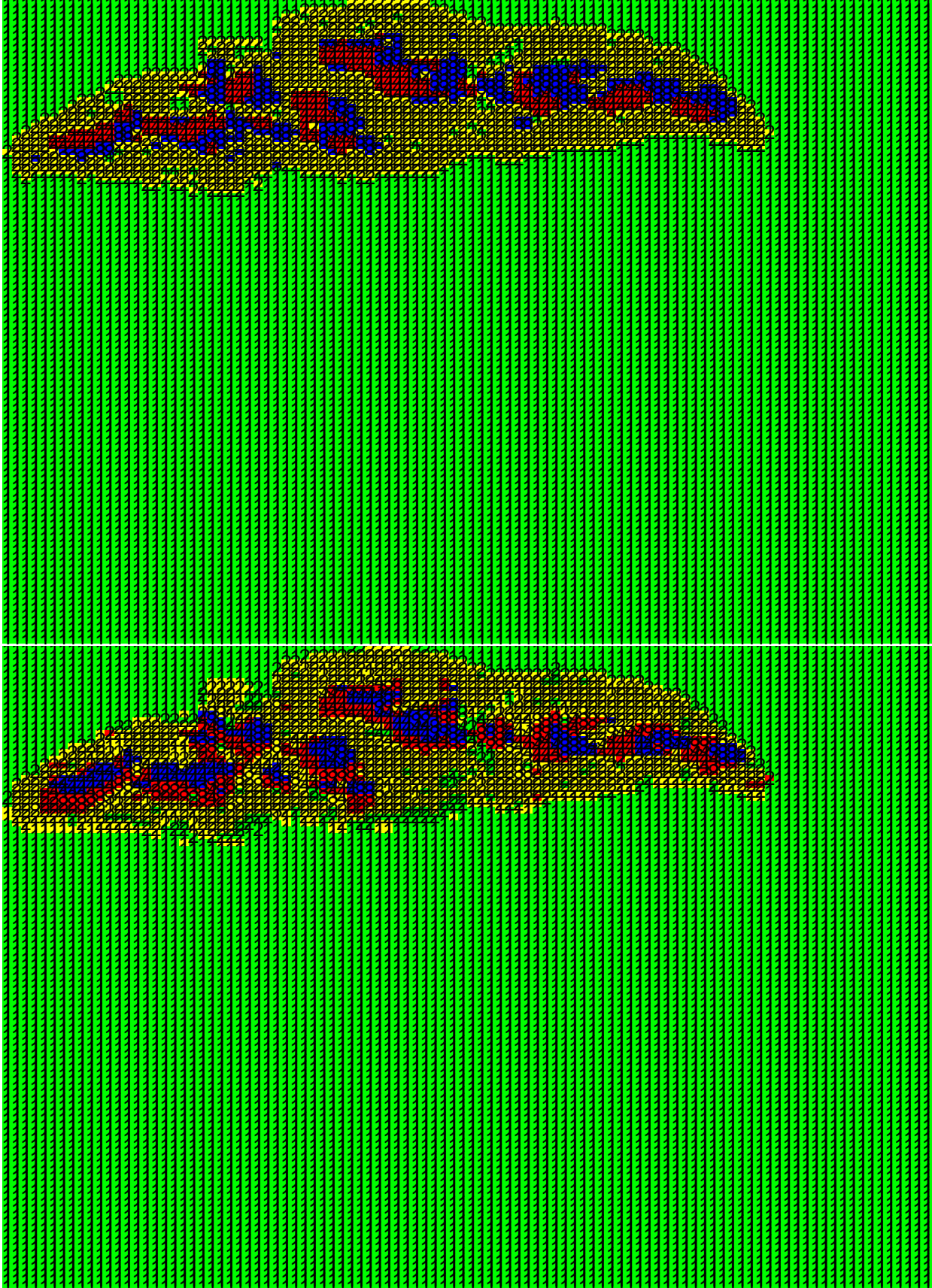
#Count1<-read.table("CD11c_1.txt",header=T,sep="\t",fill=TRUE)

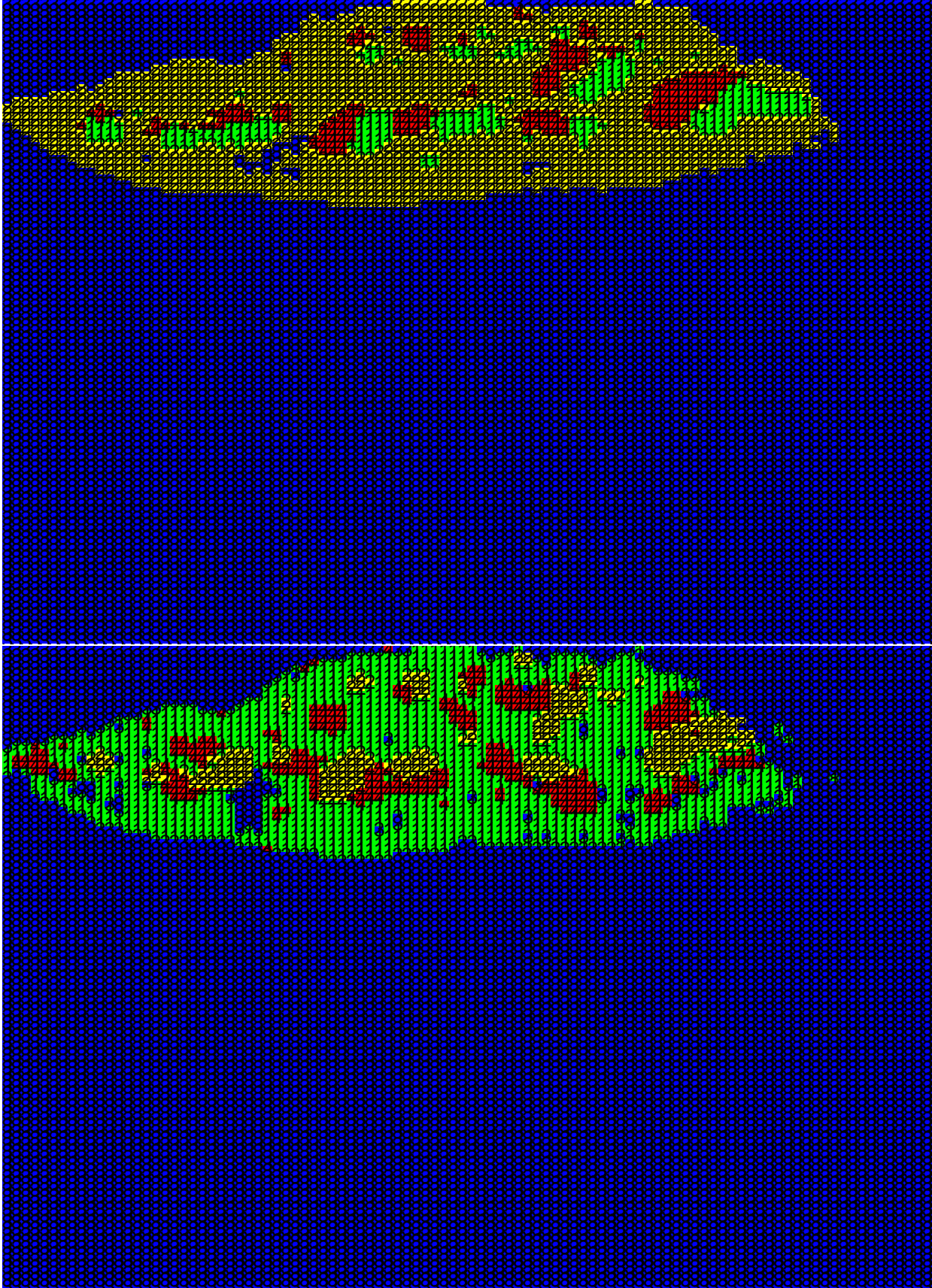
#nb the first in WT1 is not the same as Count1 so need to do better job on this (JPW)

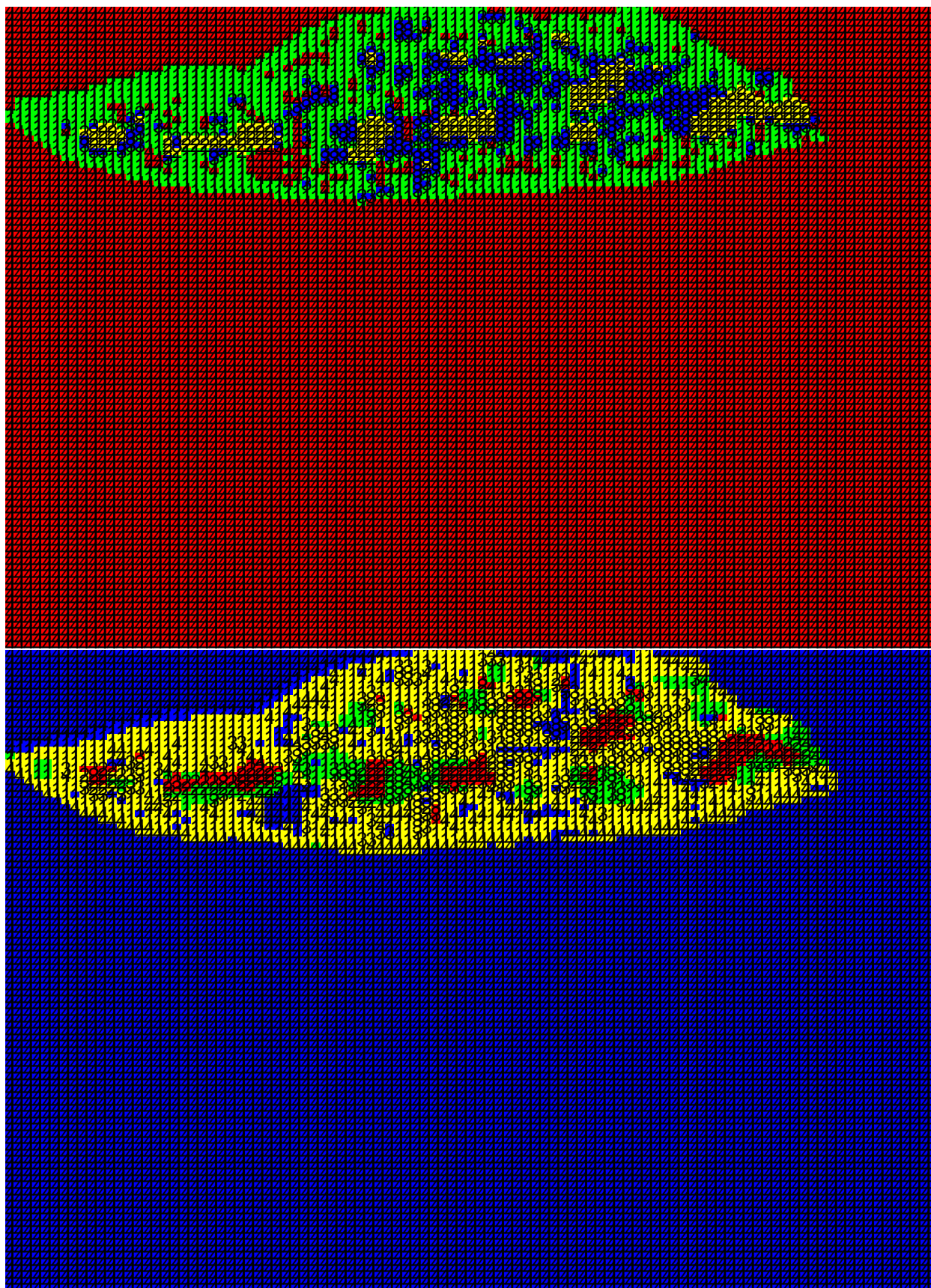
#still working on this!
# IDEALLY the variables being passed through the rest of the document are more informative and/or can be used
# e.g. name of thymus with indices for cell type and layer WT1[2,1]
```

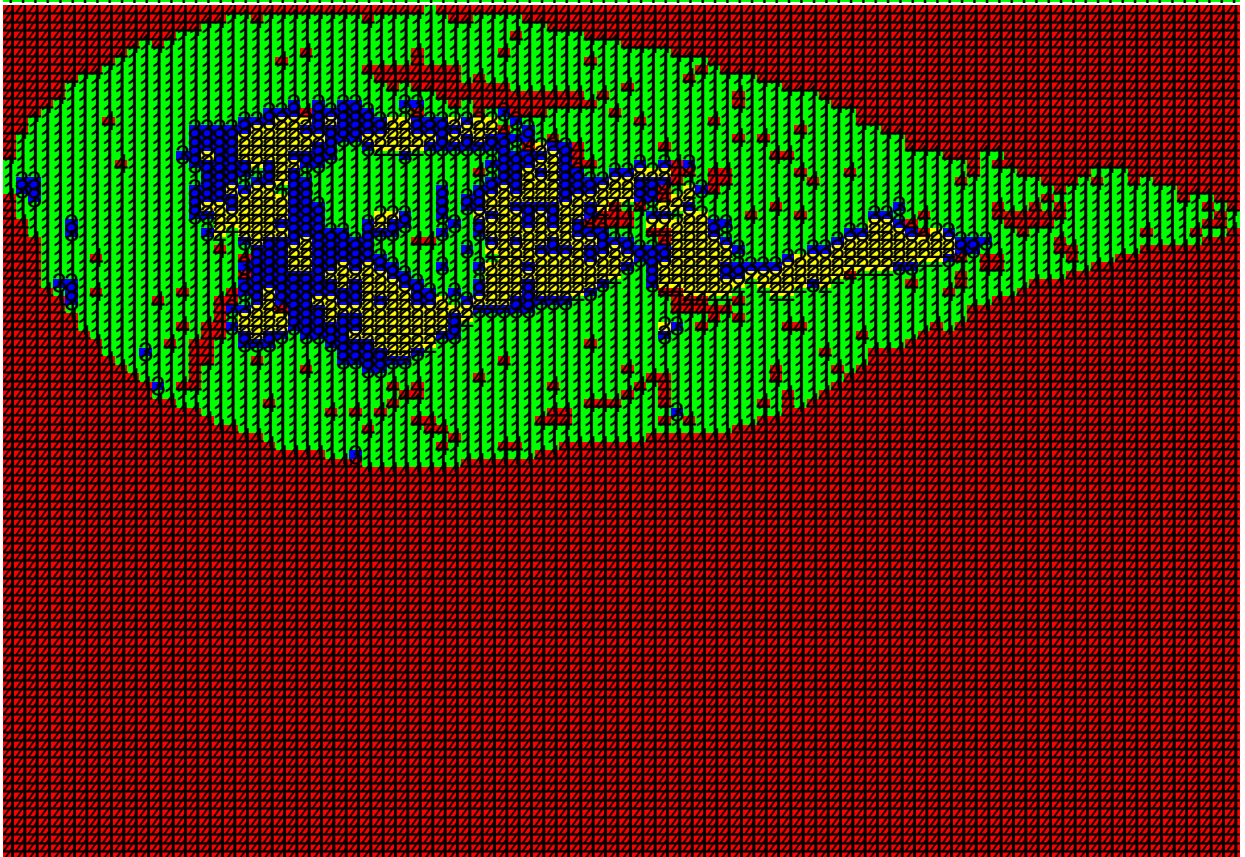
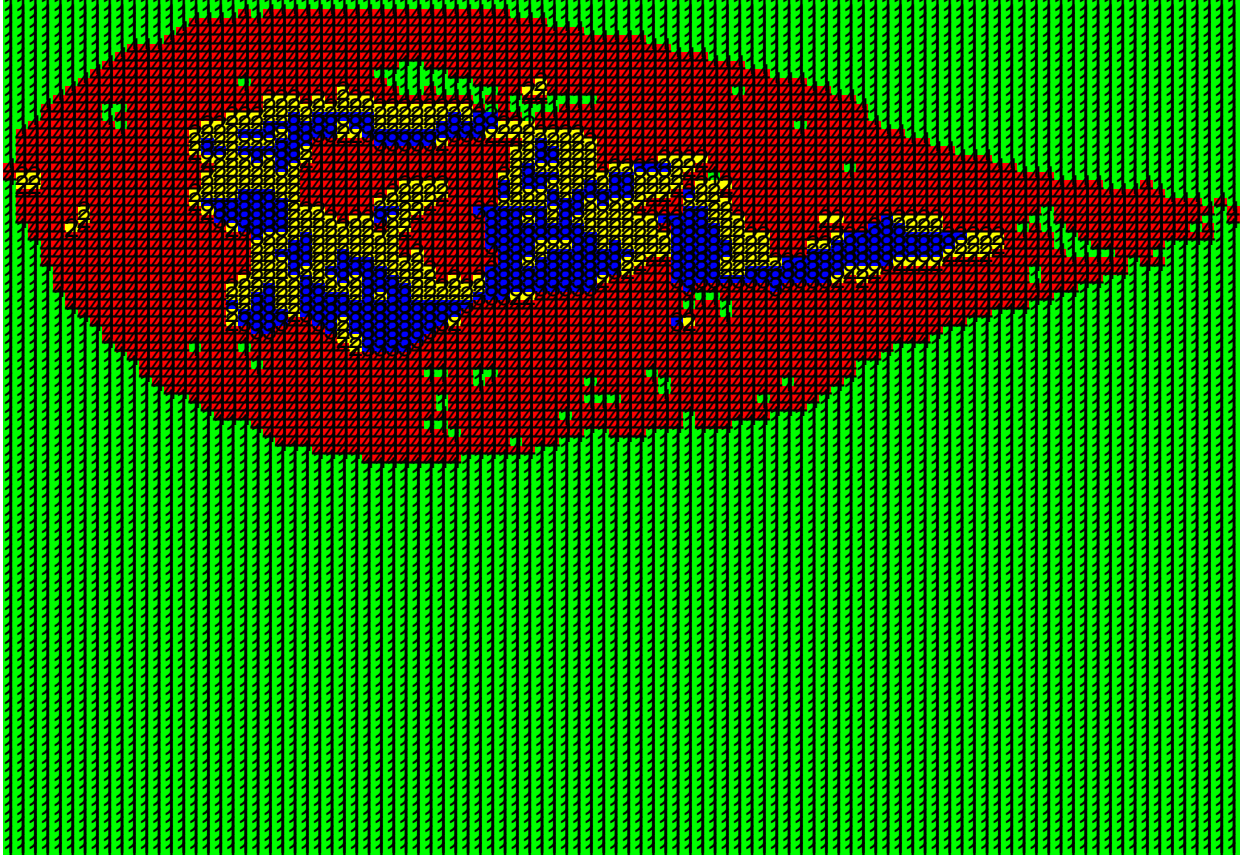
The next step is clustering. First, download and load these packages:

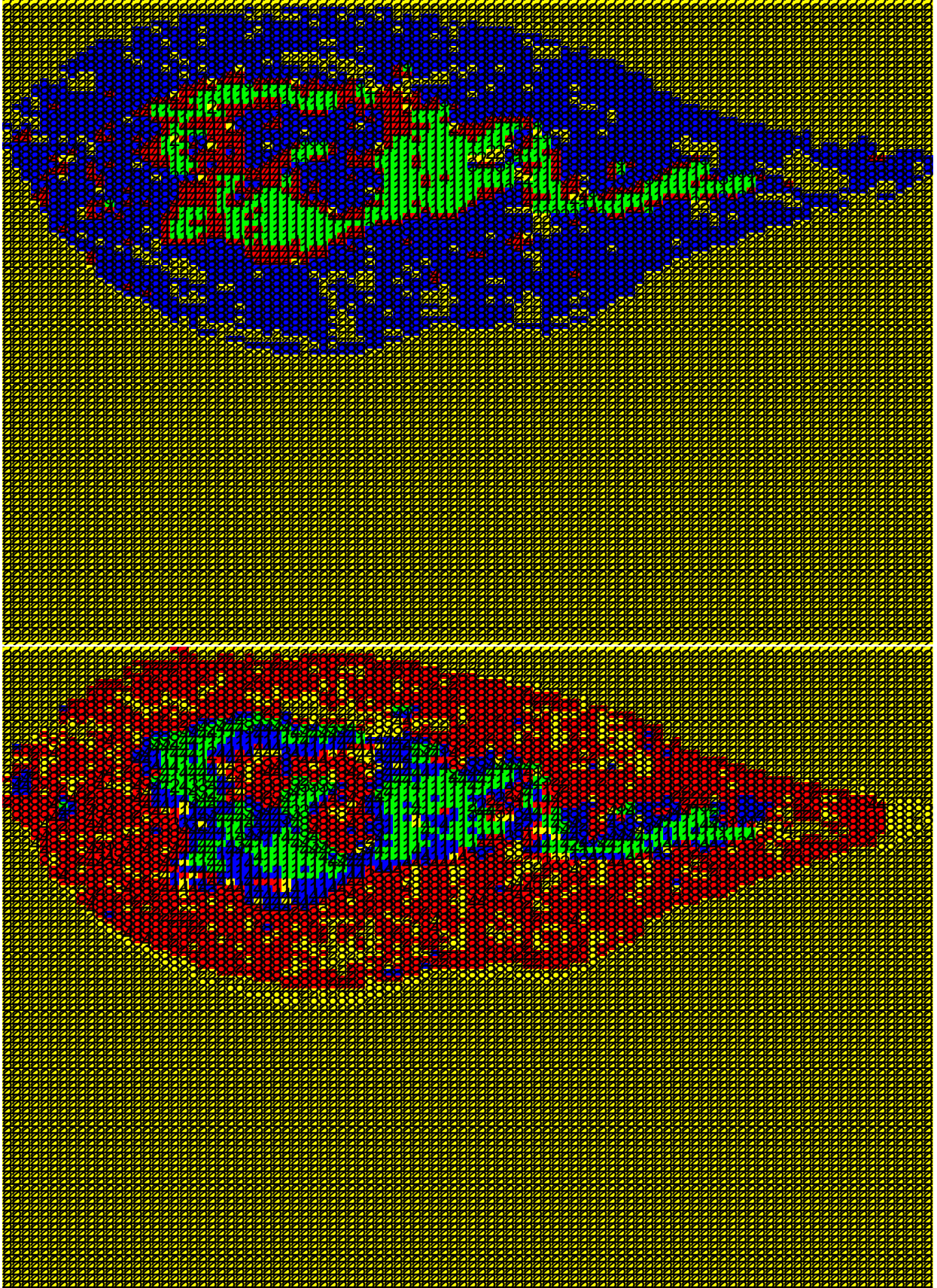






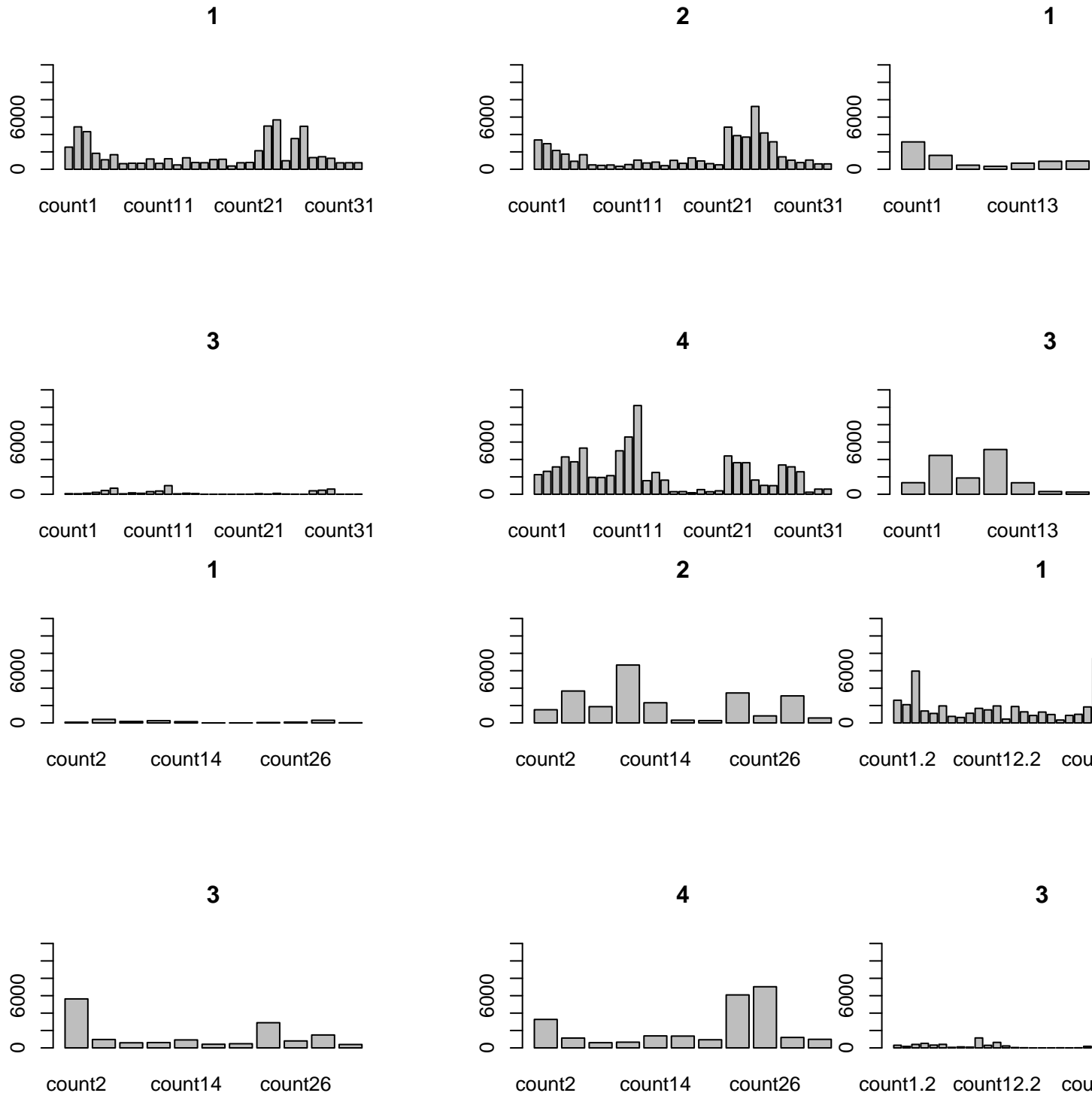


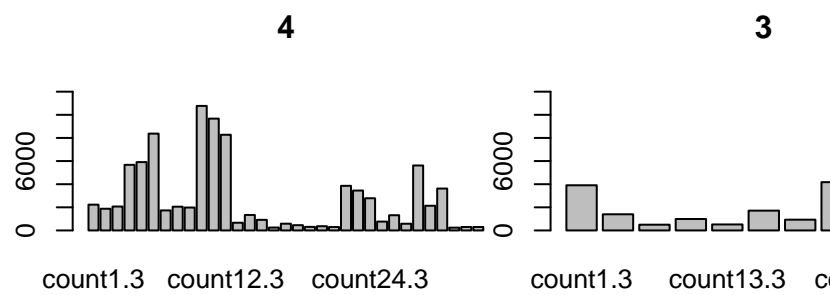
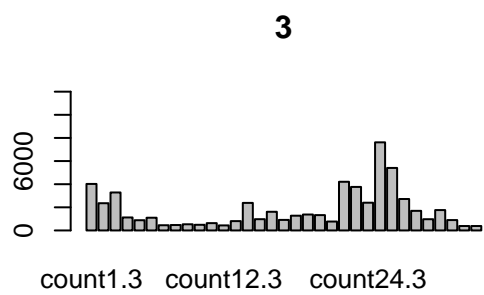
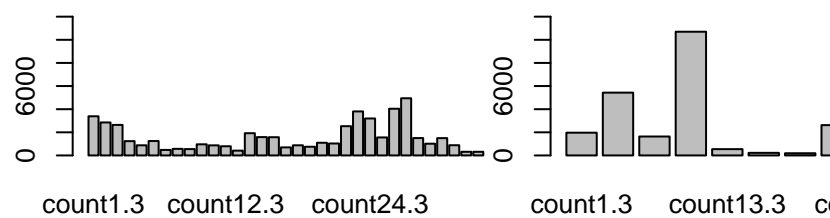
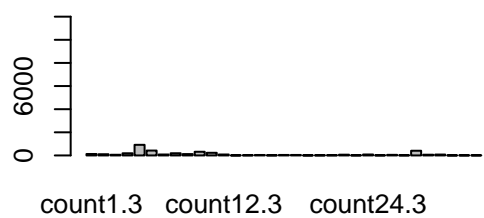
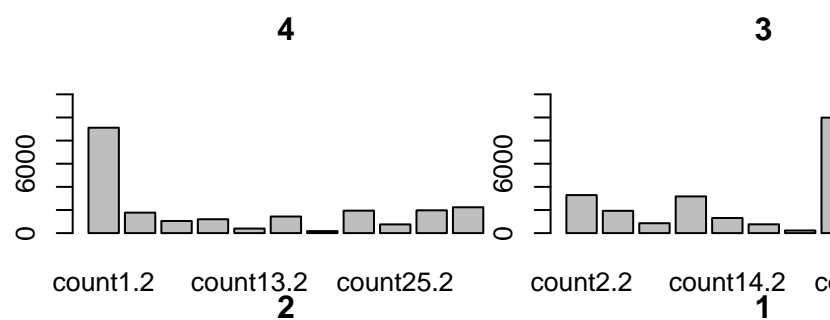
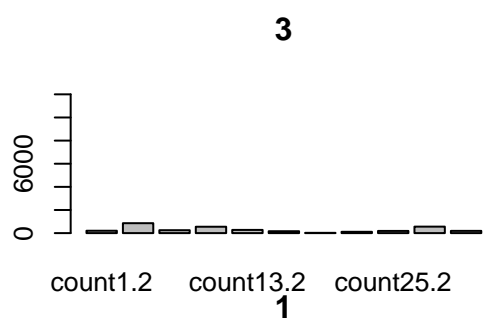
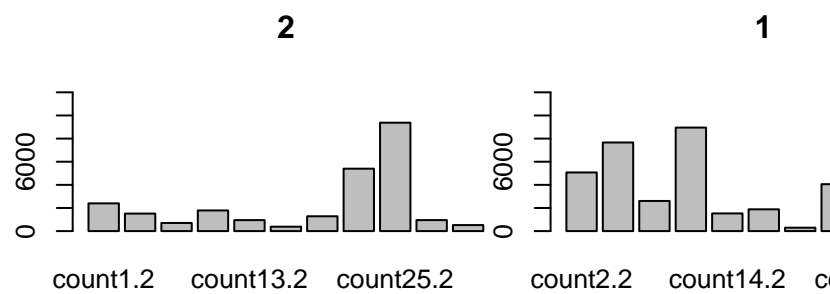
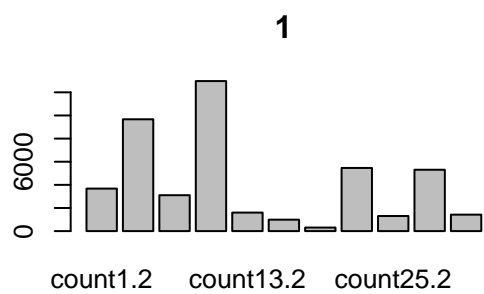


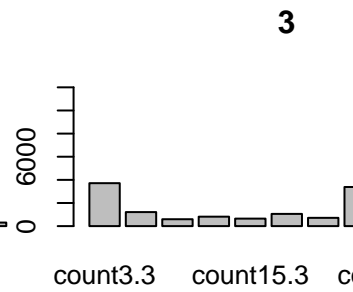
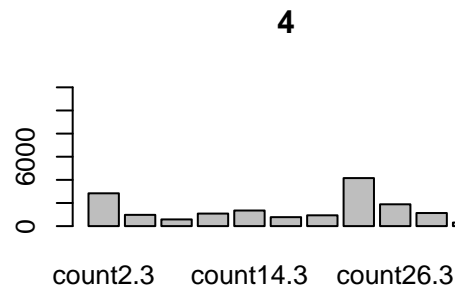
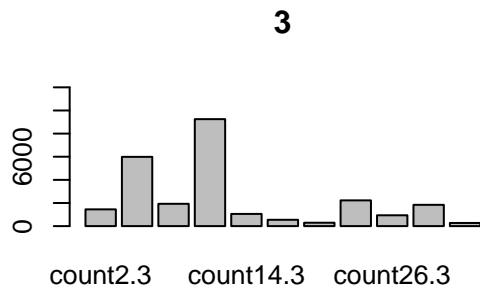
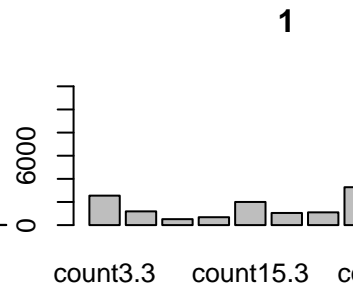
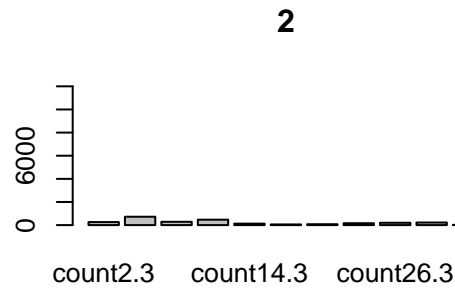
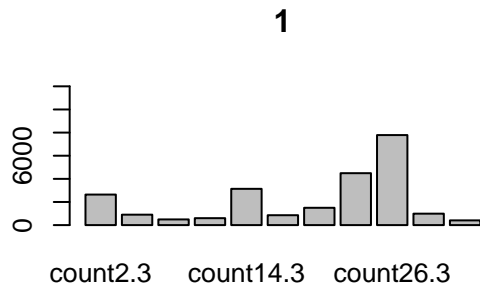


Next Step: Figuring out how clustering is done We can see that we've asked R to take cell count info from 11 cell types and make 4 clusters

It turns out that cluster 1 is just *empty* space - but not clusters 2-4 #need automatic way to identify *empty* outside of thymus



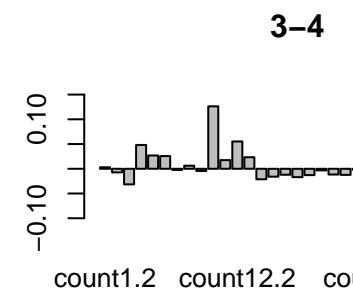
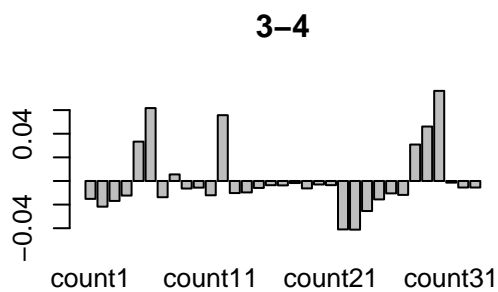
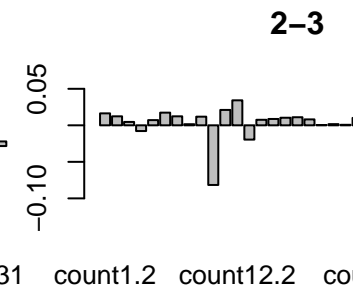
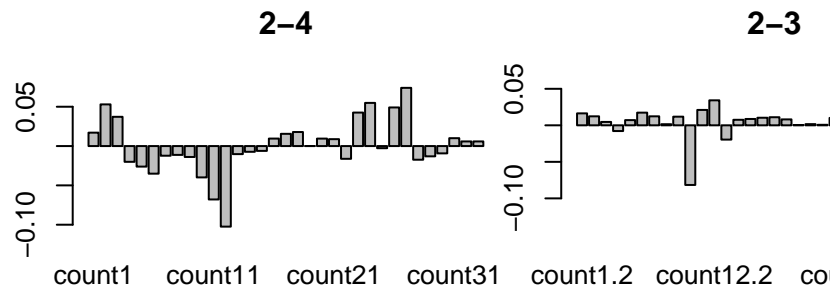
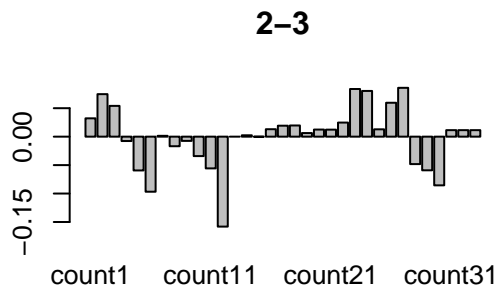




We can make a running sum of all the counts of all cell types by cluster

and normalize these so that each cluster type has the same total amount of cells (in other words, we're getting the proportion of cell types in each cluster)

We can look at the difference between pairs of clusters to see which cell types are most different between clusters



The way to read the last plots is for something like "2-3", positive bars mean that cell type is more common in cluster 2 (than 3); negative bars mean the opposite. The higher the absolute number on the y-axis, the

greater the difference is in that cell count between the two clusters. I think (as John suggested), that we can move ahead and be more formal with this analysis (e.g. Bray Curtis dissimilarity), but this is a useful first step.

Mary's plan for Bray Curtis

In R, the analogy to re-organizing data by either pivot tables, or manual calculation and pasting into a new sheet, is to manipulate one data frame to another. The 'dplyr' package or the 'aggregate' function can both help you here. For most standard ecology metrics there are usually one or more packages that make our life easier. Here the 'vegan' package is useful. In this example, we work out Bray-Curtis for all clusters (pairwise comparison)

df<-NULL df2<-NULL #these nulls are here to reset the dataframes from the rbind that is below, only important if iteratign across commands outside of KNITR

```
library(vegan) # you may need to install this if you don't have it
library(dplyr) # ditto (used for chaining and manipulation)
# http://blog.rstudio.org/2014/01/17/introducing-dplyr/
#df<-spe[which(spe.kmeans$cluster==1),] %>% colSums() %>% t() # example of chaining.
#reads as: take only cluster 1 data from 'spe', then take column sums (i.e. total cell type count)
#then transpose (swap rows/cols needed for bray curtis calc)

# each of these
df<-spe[which(spe.kmeans$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp<-spe[which(spe.kmeans$cluster==j),] %>% colSums() %>% t()
  df<-rbind(df,tmp)
}

df1.1<-spe1.1[which(spe.kmeans1.1$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp1.1<-spe1.1[which(spe.kmeans1.1$cluster==j),] %>% colSums() %>% t()
  df1.1<-rbind(df1.1,tmp1.1)
}

df1.2<-spe1.2[which(spe.kmeans1.2$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp1.2<-spe1.2[which(spe.kmeans1.2$cluster==j),] %>% colSums() %>% t()
  df1.2<-rbind(df1.2,tmp1.2)
}

df1.2<-spe1.2[which(spe.kmeans1.2$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp1.2<-spe1.2[which(spe.kmeans1.2$cluster==j),] %>% colSums() %>% t()
  df1.2<-rbind(df1.2,tmp1.2)
}

df2<-spe2[which(spe.kmeans2$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp2<-spe2[which(spe.kmeans2$cluster==j),] %>% colSums() %>% t()
  df2<-rbind(df2,tmp2)
}

df2.1<-spe2.1[which(spe.kmeans2.1$cluster==1),] %>% colSums() %>% t()
```

```

for (j in 2:4){
  tmp2.1<-spe2.1[which(spe.kmeans2.1$cluster==j),] %>% colSums() %>% t()
  df2.1<-rbind(df2.1,tmp2.1)
}

df2.2<-spe2.2[which(spe.kmeans2.2$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp2.2<-spe2.2[which(spe.kmeans2.2$cluster==j),] %>% colSums() %>% t()
  df2.2<-rbind(df2.2,tmp2.2)
}

df2.2<-spe2.2[which(spe.kmeans2.2$cluster==1),] %>% colSums() %>% t()
for (j in 2:4){
  tmp2.2<-spe2.2[which(spe.kmeans2.2$cluster==j),] %>% colSums() %>% t()
  df2.2<-rbind(df2.2,tmp2.2)
}

rownames(df)<-c("cluster1","cluster2","cluster3","cluster4")

df<-apply(df,2,as.integer) %>% as.data.frame()
df1.1<-apply(df1.1,2,as.integer) %>% as.data.frame()
df1.2<-apply(df1.2,2,as.integer) %>% as.data.frame()
df1.2<-apply(df1.2,2,as.integer) %>% as.data.frame()

df2<-apply(df2,2,as.integer) %>% as.data.frame()
df2.1<-apply(df2.1,2,as.integer) %>% as.data.frame()
df2.2<-apply(df2.2,2,as.integer) %>% as.data.frame()
df2.2<-apply(df2.2,2,as.integer) %>% as.data.frame()

#need same number of columns to bind them.
colnames(df2)<-colnames(df)

colnames(df1.1)=colnames(df2.1)
colnames(df1.2)<-colnames(df2.1)
colnames(df1.2)<-colnames(df2.1)
colnames(df2.1)<-colnames(df2.1)
colnames(df2.2)<-colnames(df2.1)
colnames(df2.2)<-colnames(df2.1)
# this is it column names were wrong

dfTOTAL<-rbind(df1.2,df2.2, df1.2, df1.2, df1.1, df1.1)

BrayCurtis<-vegdist(dfTOTAL,method="bray")
print(BrayCurtis)

```

	1	2	3	4	5	6
2	0.88148262					
3	0.81517126	0.45879273				
4	0.88191713	0.51636778	0.37024760			
5	0.93199912	0.30614749	0.51146226	0.52890043		
6	0.88065923	0.50595019	0.45764426	0.10281213	0.57772343	
7	0.88453208	0.37072828	0.37706297	0.30043512	0.45040199	0.35775784
8	0.36509240	0.76143826	0.66380289	0.77420162	0.85929563	0.75988731

9	0.00000000	0.88148262	0.81517126	0.88191713	0.93199912	0.88065923
10	0.88148262	0.00000000	0.45879273	0.51636778	0.30614749	0.50595019
11	0.81517126	0.45879273	0.00000000	0.37024760	0.51146226	0.45764426
12	0.88191713	0.51636778	0.37024760	0.00000000	0.52890043	0.10281213
13	0.00000000	0.88148262	0.81517126	0.88191713	0.93199912	0.88065923
14	0.88148262	0.00000000	0.45879273	0.51636778	0.30614749	0.50595019
15	0.81517126	0.45879273	0.00000000	0.37024760	0.51146226	0.45764426
16	0.88191713	0.51636778	0.37024760	0.00000000	0.52890043	0.10281213
17	0.86991459	0.51507505	0.36325723	0.08326576	0.54619257	0.15627012
18	0.15579323	0.90241067	0.84685561	0.90277241	0.94427108	0.90172508
19	0.86178535	0.11699877	0.47083981	0.56563435	0.35674682	0.56315836
20	0.84662698	0.39773386	0.22762375	0.24120870	0.45847325	0.31037532
21	0.86991459	0.51507505	0.36325723	0.08326576	0.54619257	0.15627012
22	0.15579323	0.90241067	0.84685561	0.90277241	0.94427108	0.90172508
23	0.86178535	0.11699877	0.47083981	0.56563435	0.35674682	0.56315836
24	0.84662698	0.39773386	0.22762375	0.24120870	0.45847325	0.31037532
	7	8	9	10	11	12

2						
3						
4						
5						
6						
7						
8	0.76719428					
9	0.88453208	0.36509240				
10	0.37072828	0.76143826	0.88148262			
11	0.37706297	0.66380289	0.81517126	0.45879273		
12	0.30043512	0.77420162	0.88191713	0.51636778	0.37024760	
13	0.88453208	0.36509240	0.00000000	0.88148262	0.81517126	0.88191713
14	0.37072828	0.76143826	0.88148262	0.00000000	0.45879273	0.51636778
15	0.37706297	0.66380289	0.81517126	0.45879273	0.00000000	0.37024760
16	0.30043512	0.77420162	0.88191713	0.51636778	0.37024760	0.00000000
17	0.35168054	0.77774298	0.86991459	0.51507505	0.36325723	0.08326576
18	0.90494885	0.45057822	0.15579323	0.90241067	0.84685561	0.90277241
19	0.40443364	0.72470910	0.86178535	0.11699877	0.47083981	0.56563435
20	0.24184646	0.69696417	0.84662698	0.39773386	0.22762375	0.24120870
21	0.35168054	0.77774298	0.86991459	0.51507505	0.36325723	0.08326576
22	0.90494885	0.45057822	0.15579323	0.90241067	0.84685561	0.90277241
23	0.40443364	0.72470910	0.86178535	0.11699877	0.47083981	0.56563435
24	0.24184646	0.69696417	0.84662698	0.39773386	0.22762375	0.24120870
	13	14	15	16	17	18

2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	0.88148262

```

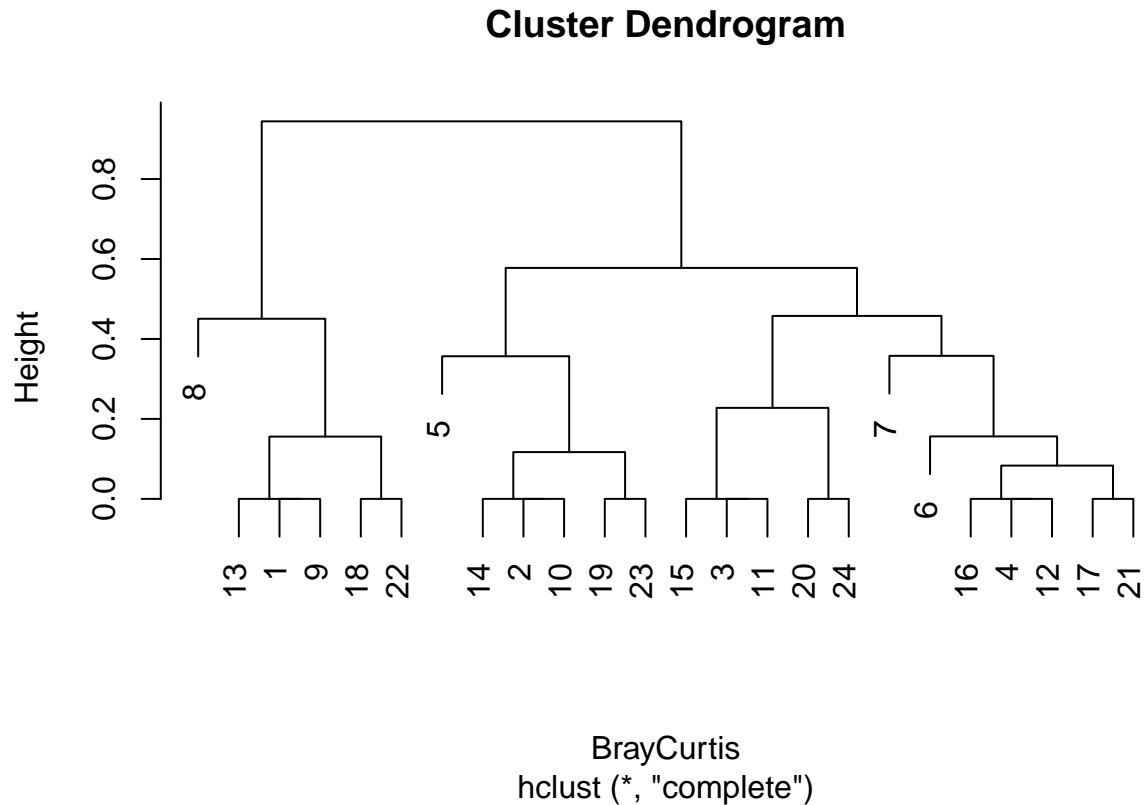
15 0.81517126 0.45879273
16 0.88191713 0.51636778 0.37024760
17 0.86991459 0.51507505 0.36325723 0.08326576
18 0.15579323 0.90241067 0.84685561 0.90277241 0.89276893
19 0.86178535 0.11699877 0.47083981 0.56563435 0.56439324 0.88598080
20 0.84662698 0.39773386 0.22762375 0.24120870 0.19695374 0.87329543
21 0.86991459 0.51507505 0.36325723 0.08326576 0.00000000 0.89276893
22 0.15579323 0.90241067 0.84685561 0.90277241 0.89276893 0.00000000
23 0.86178535 0.11699877 0.47083981 0.56563435 0.56439324 0.88598080
24 0.84662698 0.39773386 0.22762375 0.24120870 0.19695374 0.87329543
      19      20      21      22      23
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20 0.44050306
21 0.56439324 0.19695374
22 0.88598080 0.87329543 0.89276893
23 0.00000000 0.44050306 0.56439324 0.88598080
24 0.44050306 0.00000000 0.19695374 0.87329543 0.44050306

```

```

hc<-hclust(BrayCurtis)
plot(hc,labels=dfTOTAL$rownames)

```

Results

1. How many mice of each type, how many replicates of mouse x thymus x prep method.
2. Size of quadrat selected, and what this says about number of individual quadrats examined, number of total cell counts. People love those sorts of numbers!
3. Clustering and K selection
4. Using BC to validate how different clusters are, importance of composite versus individual layers.
5. Geography: Sornborger
6. etc.

Discussion

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.

1. I can generate a map of the thymus that supports previous discussion of the geography of the thymus, but it is automatic and dependent on objective criteria of sorting the functional cell types in a spatially explicit way.
2. We believe we have identified subregions in the medulla. This is cool.
3. Though our replication is minimal, we can say that there is/is not quantifiable distinction between the spatial organization of WT and Aire- mice. This has been debated in the literature and we resolve it.
4. There is much more we can do with this approach.

5. Ecologically, it is unusual to have the problem of such complete sampling of an ecosystem. Our approach to sampling this has strengths and concerns, generally associated with K-means clustering which has its subjective/ambiguous problems but is the best we can do!

Literature Cited

Angelini, C., A. H. Altieri, B. R. Silliman, and M. D. Bertness. 2011. "Interactions Among Foundation Species and Their Consequences for Community Organization, Biodiversity, and Conservation." Journal Article. *Bioscience* 61 (10): 782–89. doi:Doi 10.1525/Bio.2011.61.10.8. \T1\textless{}\GotoISI\T1\textgreater{}://WOS:000299805400008.

Verberk, W. 2011. "Explaining General Patterns in Species Abundance and Distributions." Journal Article. *Nature Education Knowledge* 3: 38.