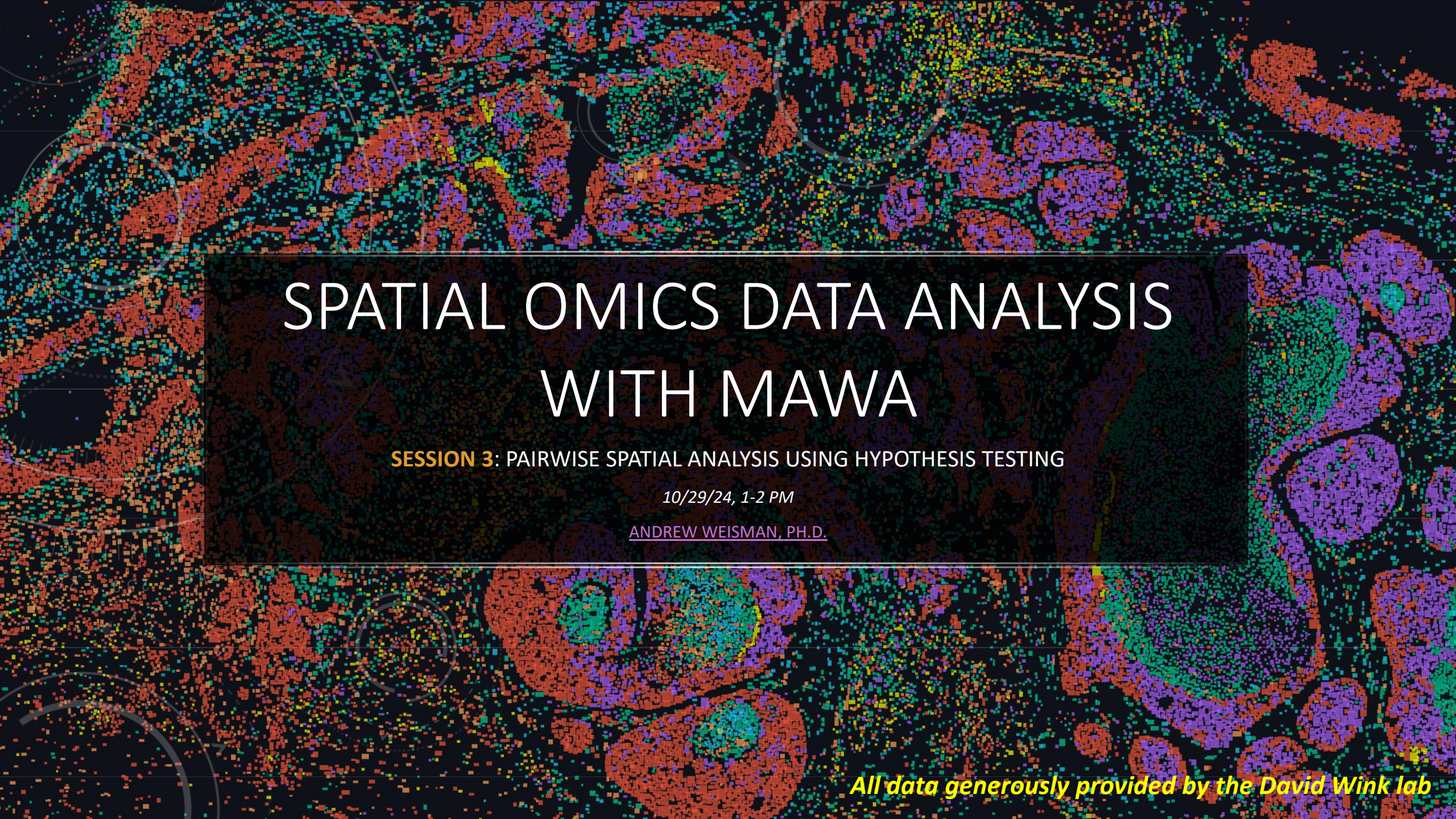


SPATIAL OMICS DATA ANALYSIS WITH MAWA

SESSION 3: PAIRWISE SPATIAL ANALYSIS USING HYPOTHESIS TESTING

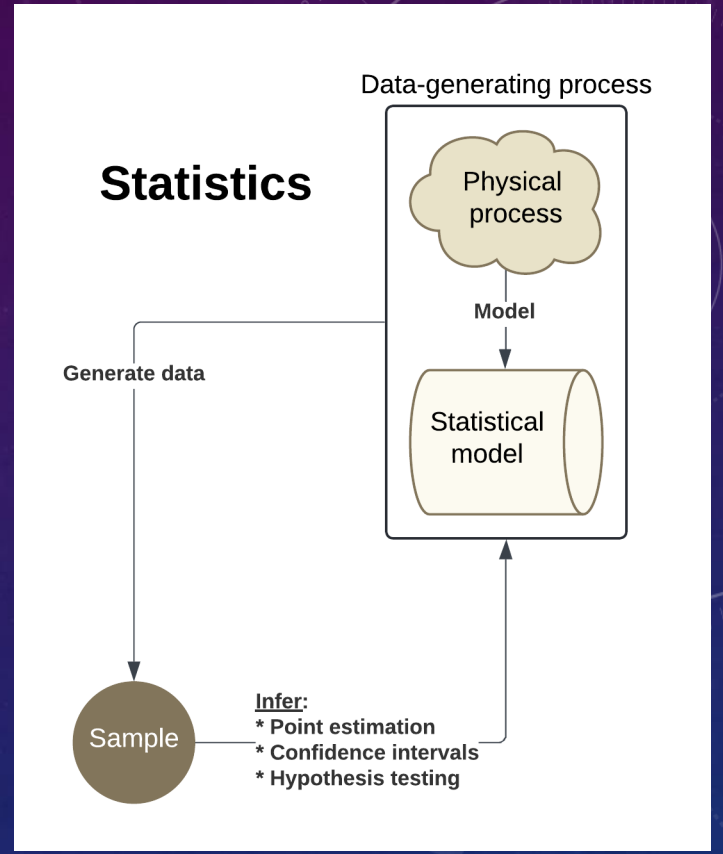
10/29/24, 1-2 PM

ANDREW WEISMAN, PH.D.

All data generously provided by the David Wink lab

STATISTICS OVERVIEW

- **What is a statistic?** A function of a sample of data. *E.g., mean, maximum, standard deviation, correlation, chance of rain, defect rate, T statistic.*
- **Two thrusts of statistics**:
    1. **Descriptive statistics**: Calculate statistics for the point of describing a sample of data. *E.g., what is the median value of the sample?*
    2. **Inferential statistics**: Use statistics to infer properties of the process that generated the data. *E.g., based on the sample data, what is the mean of the "population distribution" that generated the data?*
- **Common problems in (inferential) statistics**:
    1. **(Point) estimation**: Obtain a single "best guess" of a quantity of interest. *E.g., given 10 coin flips, what's the probability of obtaining a head? I.e., what's the estimate of the parameter p of the binomial distribution that generated the data? E.g., 0.6.*
    2. **Confidence intervals**: Obtain bounds of a quantity of interest to within a specified coverage. *E.g., what is a 95% confidence interval on the estimate for p? E.g., [0.5, 0.7].*
    3. **Hypothesis testing**: Can we reject a "null hypothesis" in favor of an "alternative hypothesis"? *E.g., is the coin fair?*

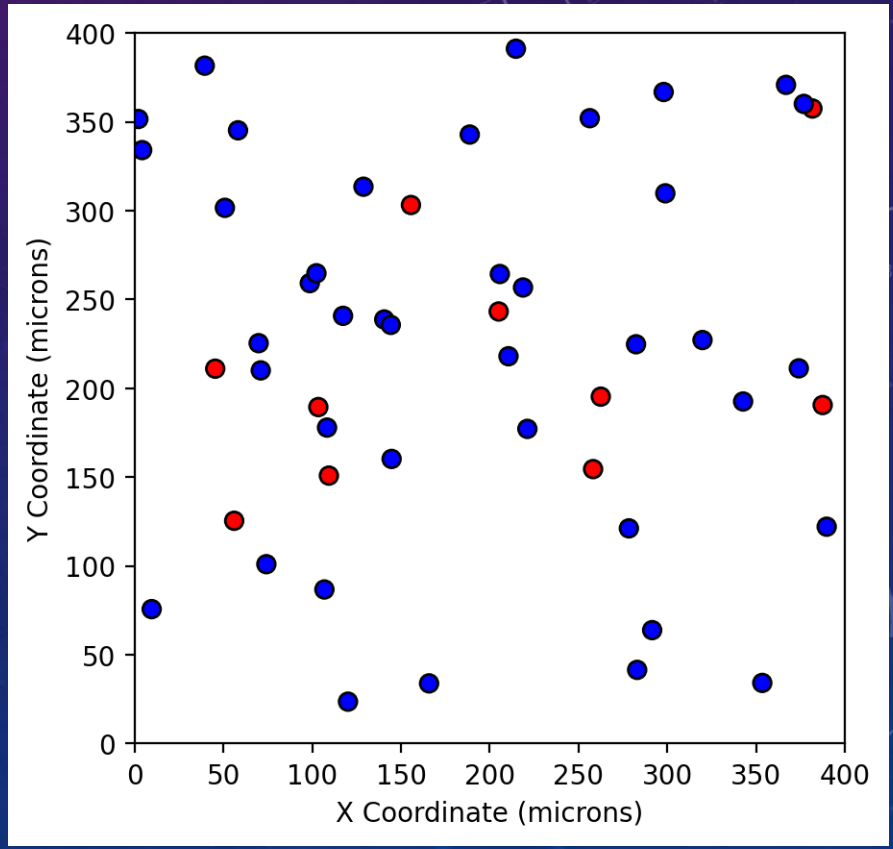


OVERVIEW OF HYPOTHESIS TESTING

- **Hypothesis testing answers questions like**:
  - **Is a population parameter a specified value?** *E.g., t-test, Wald test, chi-squared test, likelihood ratio test.*
  - **Do the data fit a particular probability distribution?** *E.g., chi-squared test, Kolmogorov-Smirnov test, Shapiro-Wilk test.*
  - **Are two statistical models the same?** *E.g., permutation test, Kolmogorov-Smirnov test, Anderson-Darling test.*
- **Hypothesis testing steps**:
  1. **Define a null hypothesis H0**. *E.g., A drug has no effect.*
  2. **Define an alternative hypothesis H1**. *E.g., The drug has an effect.*
  3. **Choose or formulate an appropriate hypothesis test for testing these hypotheses**. There is no need to choose a pre-existing test; all that is needed is to choose appropriate hypotheses and a test statistic whose distribution under H0 is known.
  4. **Use the data to calculate**: (1) the test statistic *T*, (2) the "null distribution", and (3) the resulting P value.
  5. **Make a decision using the P value**. Generally, small P values favor H1 and large P values favor H0.

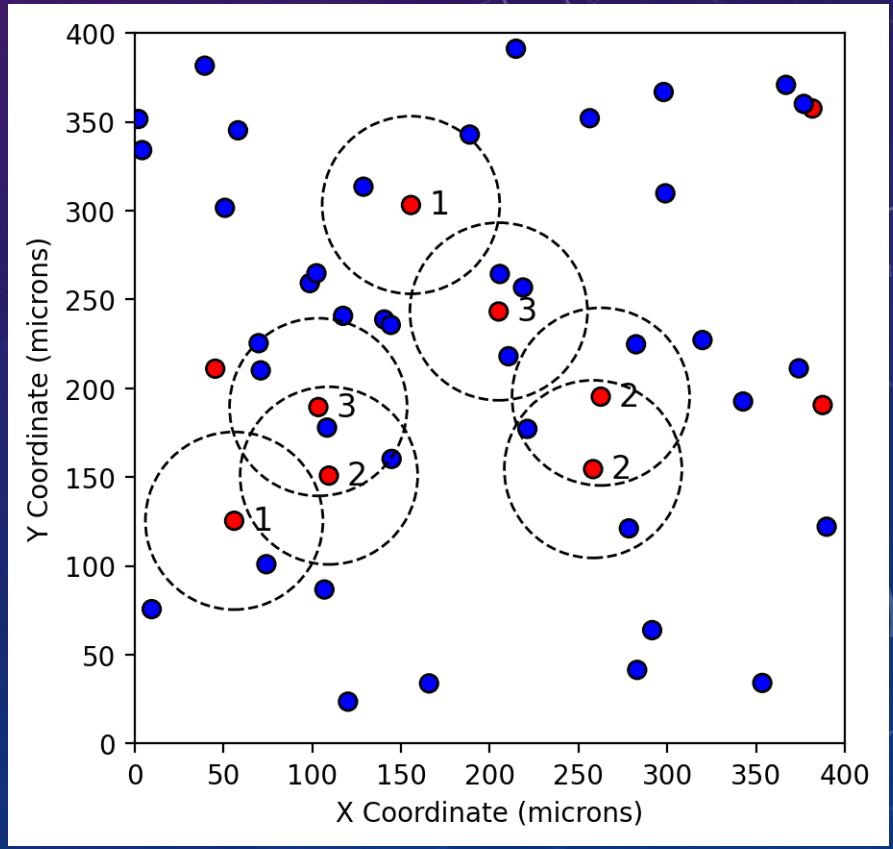
REPRESENTATIVE EXPERIMENTAL SETUP

- Plot cells as circles located at their centroids.
- There are two phenotypes of cells: red and blue.
- **Question**: What is the interaction between blue and red cells within a given “analysis radius” of 50 µm?
- Start by considering the arrangement of blue “neighbor” cells around red “center” cells.
- **Possibilities**:
  - Aggregation, i.e., clustering
  - Dispersion, i.e., repulsion
  - No particular interaction



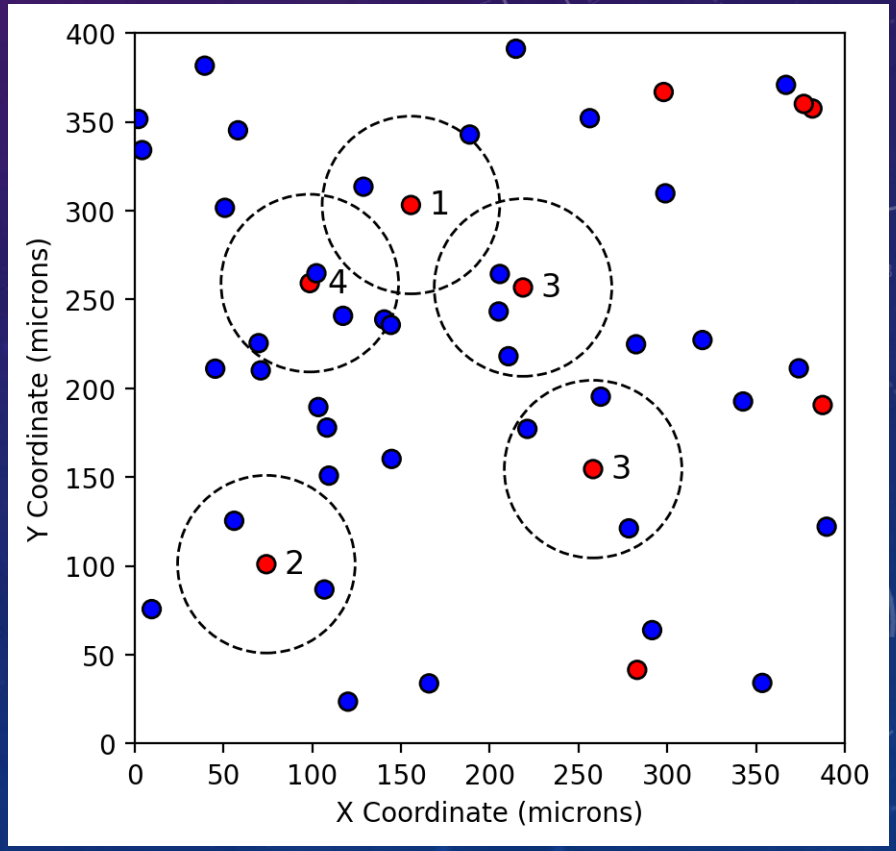
REPRESENTATIVE EXPERIMENTAL SETUP CONT.

- Draw a circle of radius 50 µm around every red center.
- Ignore centers within 50 µm of the region of interest (ROI) edge.
- Count the number of blue neighbors inside each circle and assign this neighbor count to the corresponding red center.
- Add up the counts for all red centers in the ROI.
- This defines the statistic of interest: T.
- In the example at right, the **obs**erved statistic is $T_{\text{obs}} = 14$.



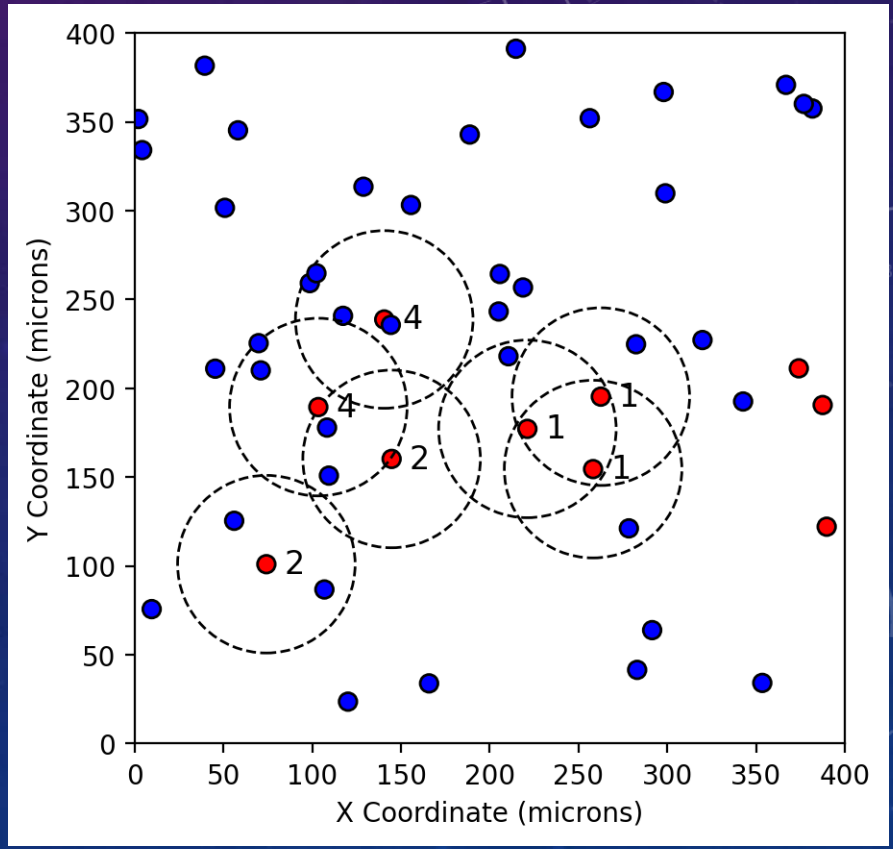
PERMUTATION METHOD

- This method employs the permutation test.
- It is a very common test backing the following type of spatial analysis.
- There are variations; what follows is the overall idea.
- Keeping the cell coordinates fixed, randomly permute the labels, i.e., phenotypes.
- Repeat the neighbor-counting experiment to obtain T_1^*.
- In the example at right, $T_1^* = 13$.



PERMUTATION METHOD CONT.

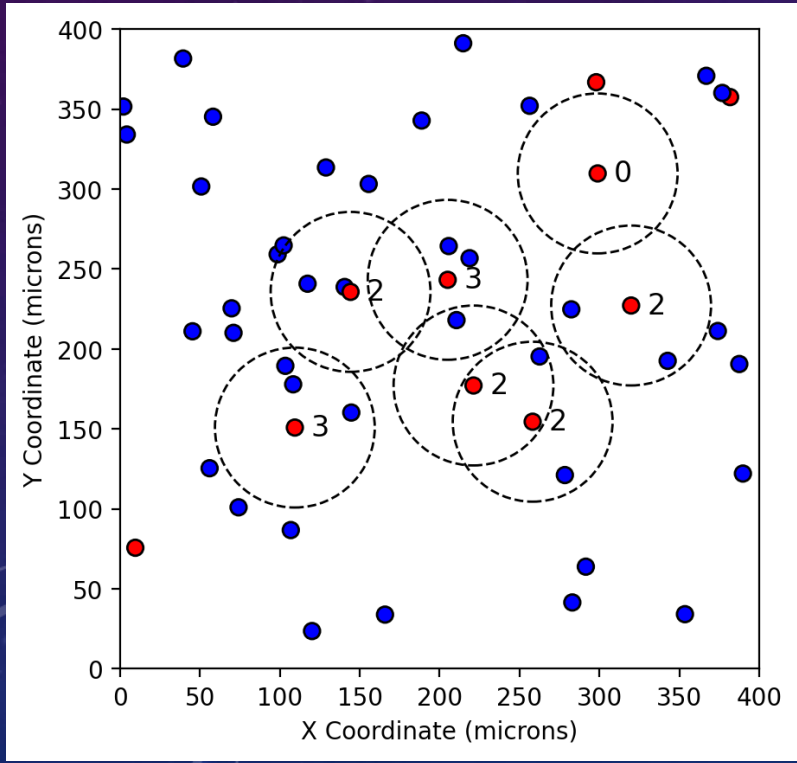
- Permute the phenotypes again to obtain T_2^*.
- In the example at right, $T_2^* = 15$.



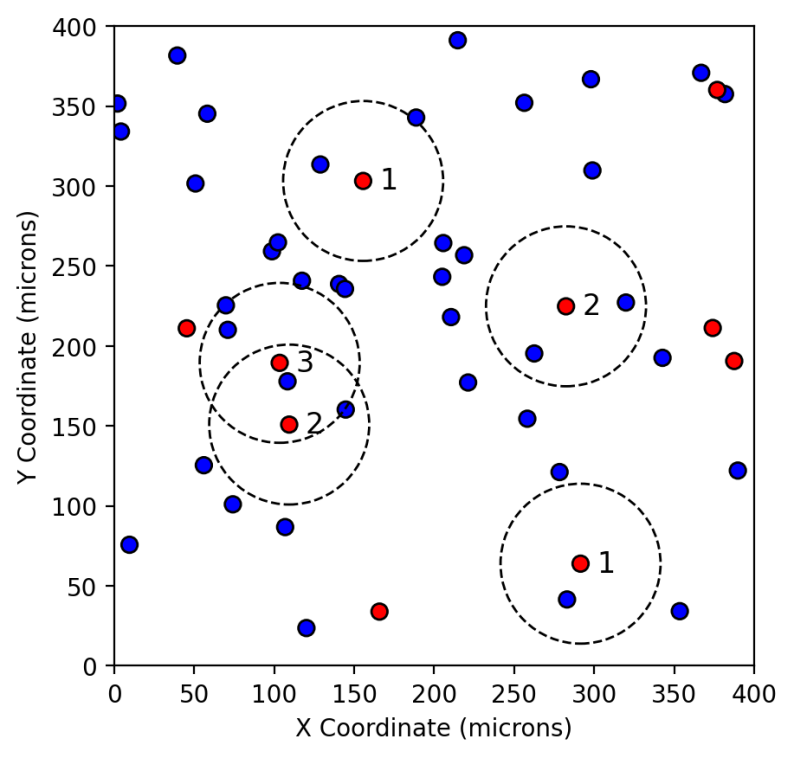
PERMUTATION METHOD CONT.

- Do this three more times:

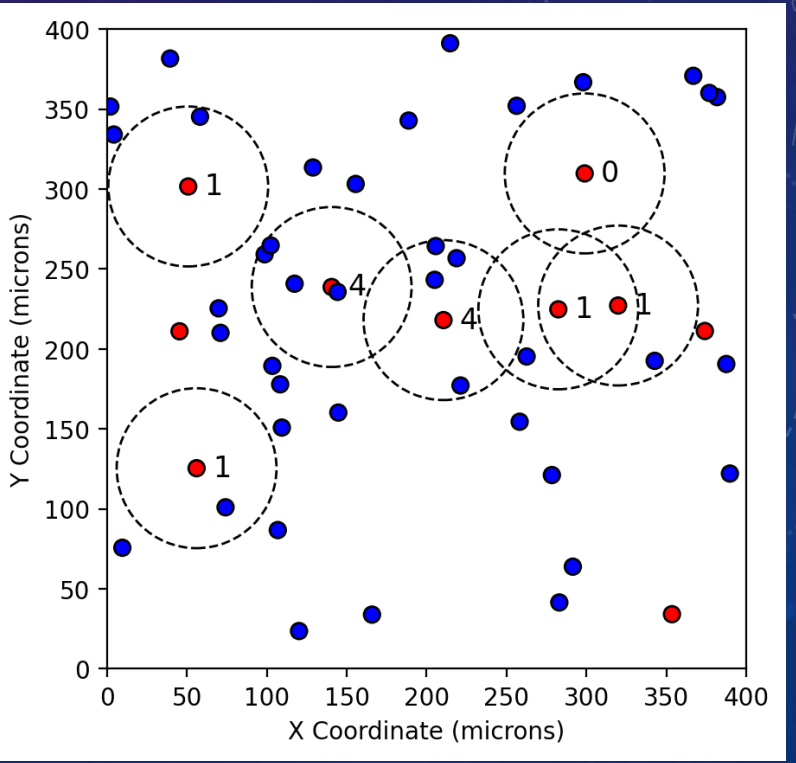
$T_3^* = 14$



$T_4^* = 9$

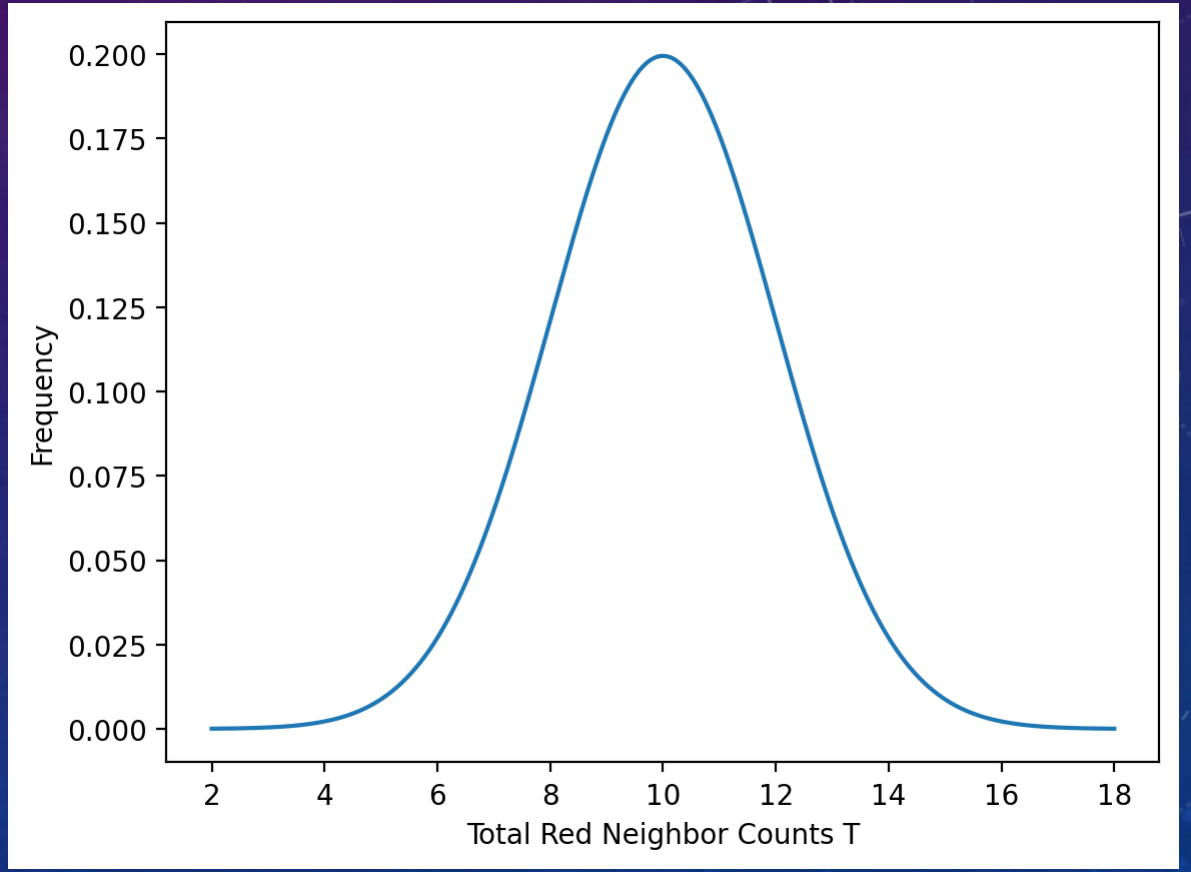


$T_5^* = 12$



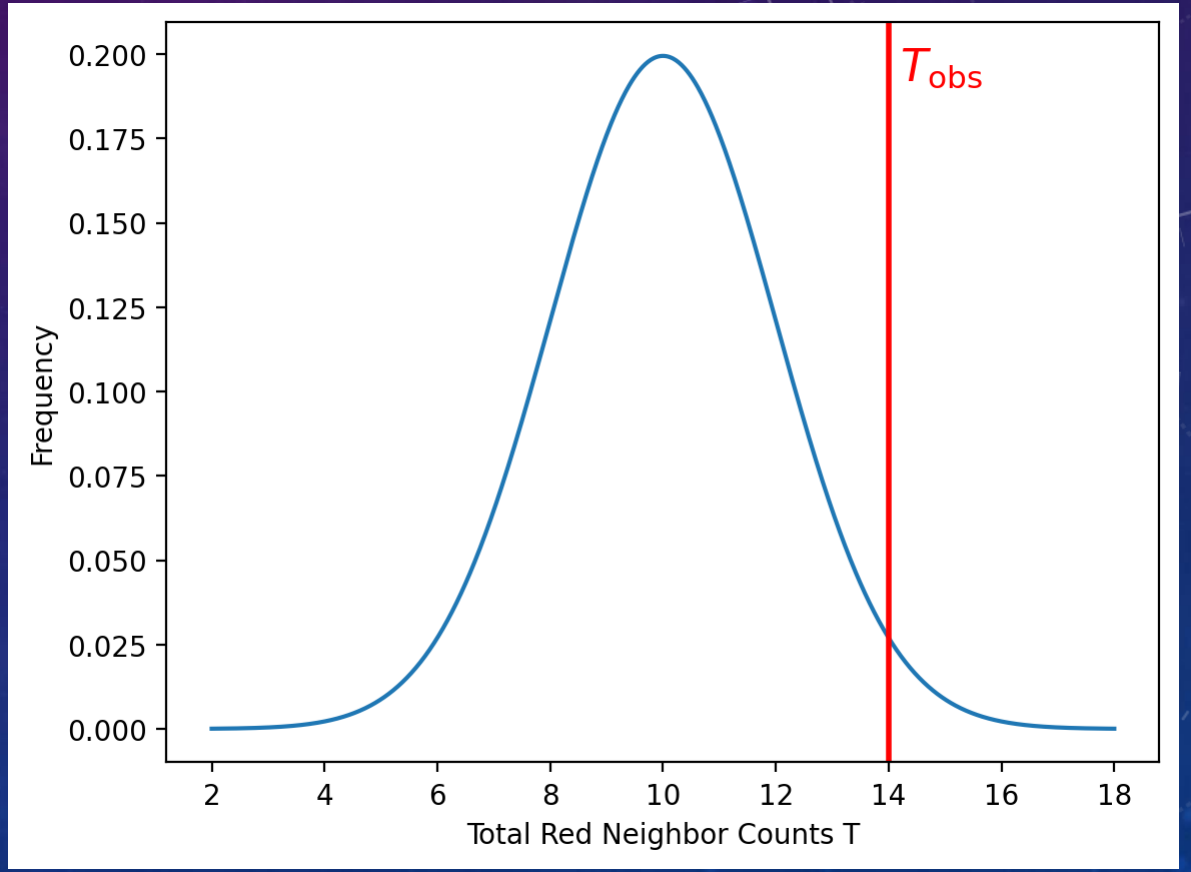
PERMUTATION METHOD CONT.

- Repeat this a total of 1000 times to obtain 1000 values of T_i^*.
- Plot the distribution of the T_i^* (i.e., make a histogram).
- **Side note**:
  - For small sample sizes (i.e., few centers), this curve will not be Gaussian.
  - For larger sample sizes (at least 10-30 centers), this curve will approximate a Gaussian due to the central limit theorem.
- This is the "null distribution," i.e., the distribution of the statistic *T* under the null hypothesis H0.
- **The permutation test in this scenario**:
  - **Null hypothesis H0**: The spatial distribution of the blue and the red cells is the same.
  - **Alternative hypothesis H1**: The distributions are different.
  - **Statistic *T***: Total number of blue neighbors around red centers.



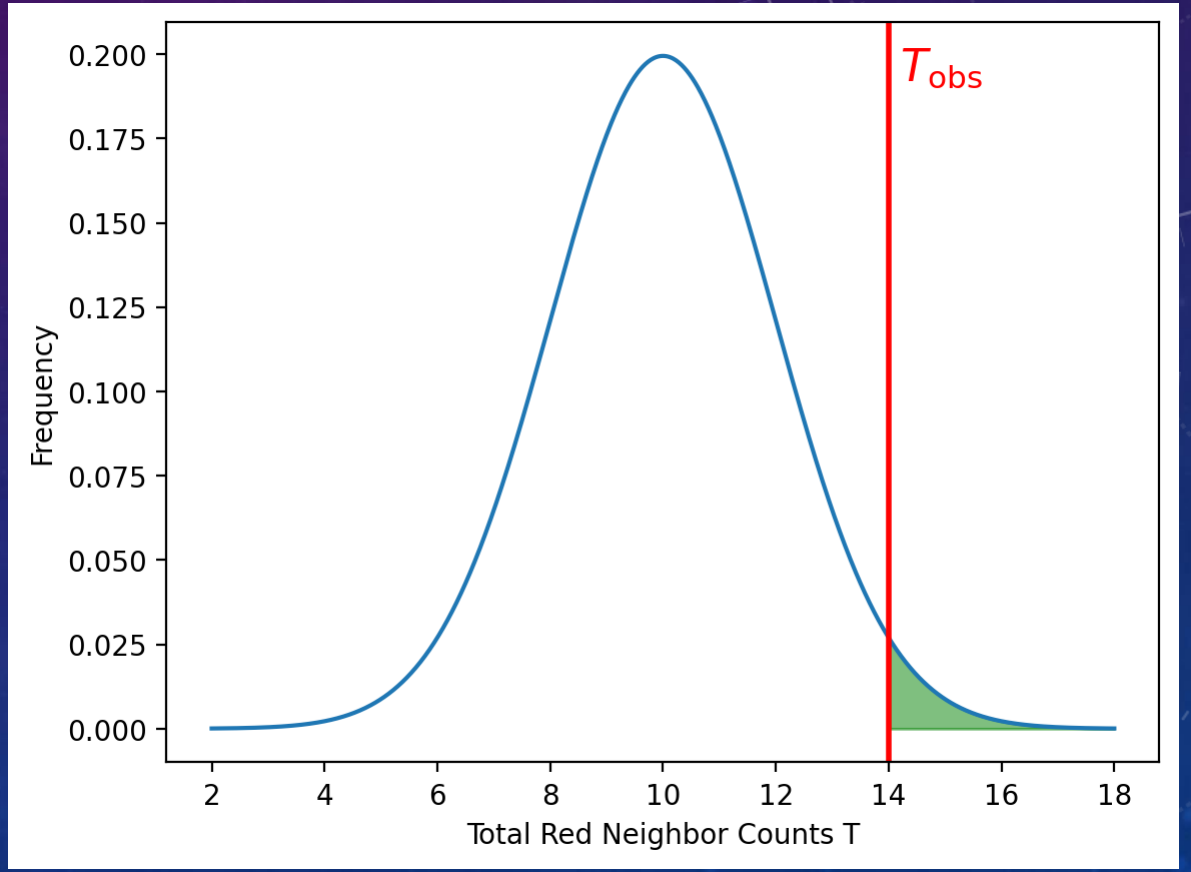
SIGNIFICANCE CALCULATION

- Overplot the observed value of the statistic *T*.



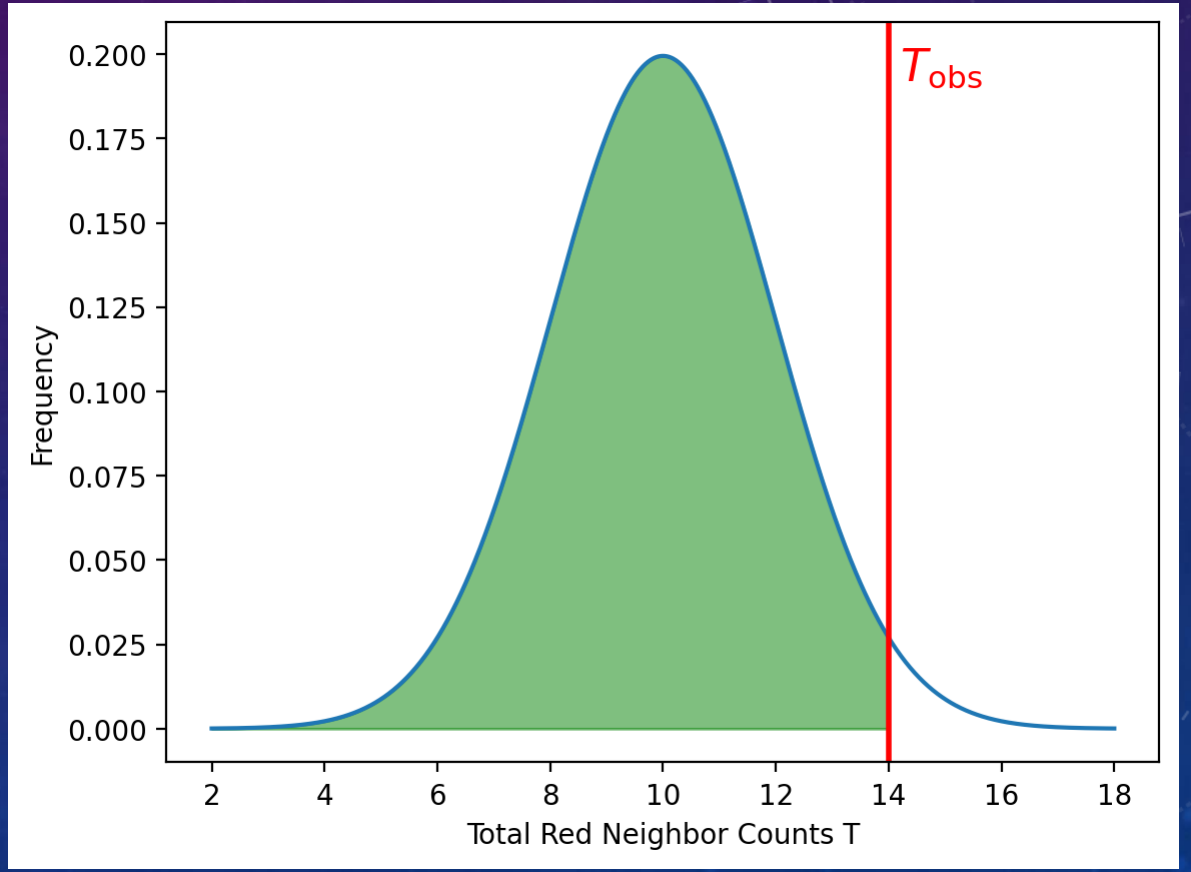
SIGNIFICANCE CALCULATION CONT.

- **Right P value**: Calculate the area under the curve to the **right** of T_{obs}:
  - $P_{\text{right}} = 0.023.$
- **Calculate the logarithm of the P value**:
  - $\log_{10} P_{\text{right}} = -1.64.$
- **Small right** P value implies that T_{obs} is **greater** than what it would be under the null hypothesis.
- **Ranges**:
  - ***P***: $[0, 1]$ *(always positive, the smaller the more significant)*
  - **$\log_{10} P$**: $[-\infty, 0]$ *(always negative, the smaller the more significant)*
- **P value interpretation**:
  - **Small right P value**: Observed statistic is "greater" than expected.
  - **Small left P value**: Observed statistic is "less" than expected.
  - **Heatmap colorbar**: The smaller the P value, the darker the color. I.e., the darker the color, the more significant the result.



SIGNIFICANCE CALCULATION CONT.

- **Left P value**: Calculate the area under the curve to the **left** of T_{obs}:
  - $P_{\text{left}} = 0.977.$
- **Calculate the logarithm of the P value**:
  - $\log_{10} P_{\text{left}} = -0.01.$
- **Large left** P value implies that T_{obs} is **not less** than what it would be under the null hypothesis.

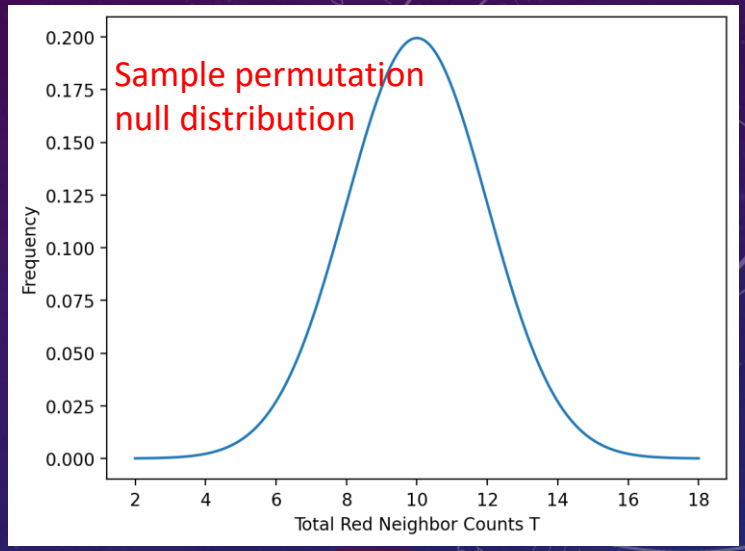


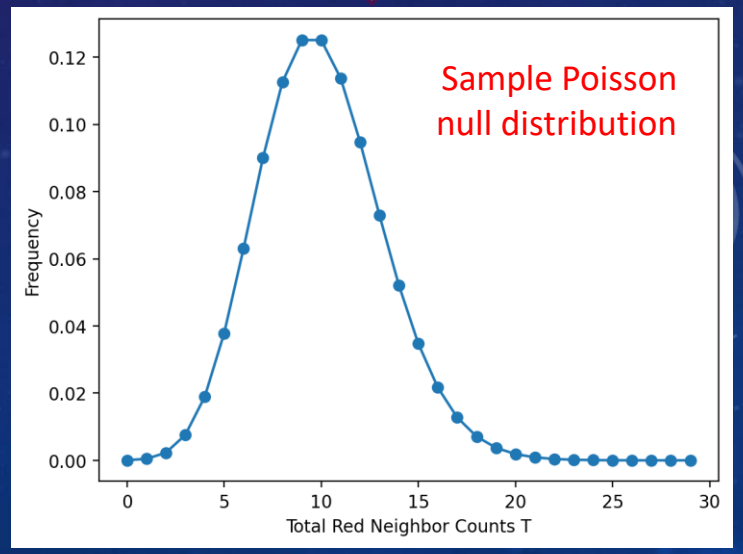
PERMUTATION METHOD INTERPRETATION

- If a P value is small, we reject the null hypothesis H0 that the spatial distribution of the blue and the red cells is the same **with respect to the statistic *T***.
- Instead, we favor the alternative hypothesis H1 that the spatial distribution is different **with respect to the statistic *T***.
- With respect to our chosen statistic *T* (total number of blue neighbors around red centers):
  - A small right P value ***often*** implies aggregation/clustering of the blue and red cells.
  - A small left P value ***often*** implies dispersion/repulsion of the blue and red cells.
- However, this is not ***always*** true. The results of the permutation method applied this way can lead to misleading conclusions.

ALTERNATIVE: POISSON METHOD

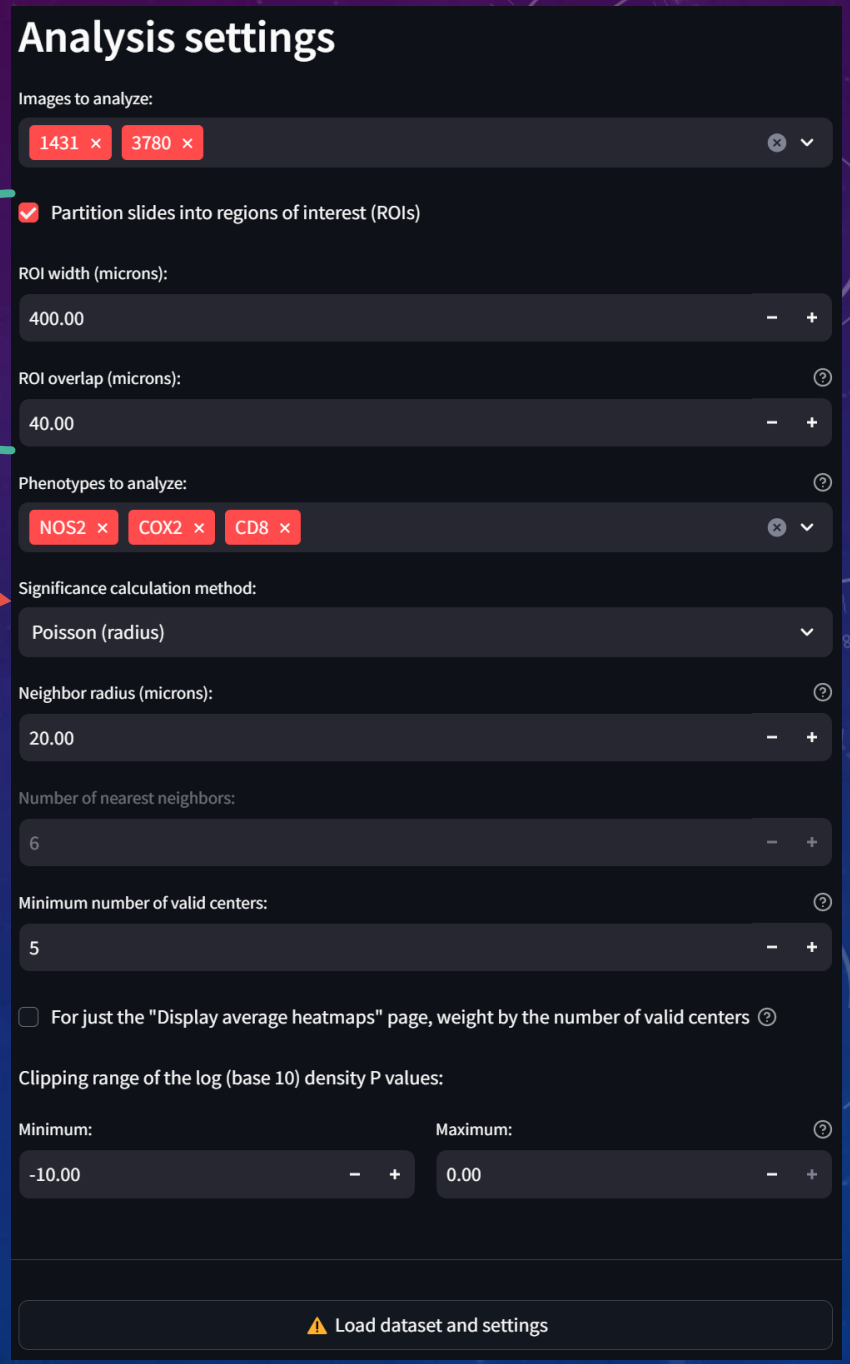
- **Define a new hypothesis test**:
  - **Null hypothesis H0**: The spatial distribution of blue neighbors is completely random.
  - **Alternative hypothesis H1**: The distribution is not completely random.
  - **Statistic *T*:** Total number of blue neighbors around red centers (same as before).
- **Procedurally, there is only a single difference from the permutation test**: Use a different distribution of *T* under the null hypothesis H0.
  - We know that under H0, *T* follows a Poisson distribution.
- **Interpretation**:
  - A small right P value ***necessarily*** implies aggregation/clustering of the blue and red cells.
  - A small left P value ***necessarily*** implies dispersion/repulsion of the blue and red cells.
- The results from the Poisson method will ***necessarily*** lead to physically intuitive conclusions.



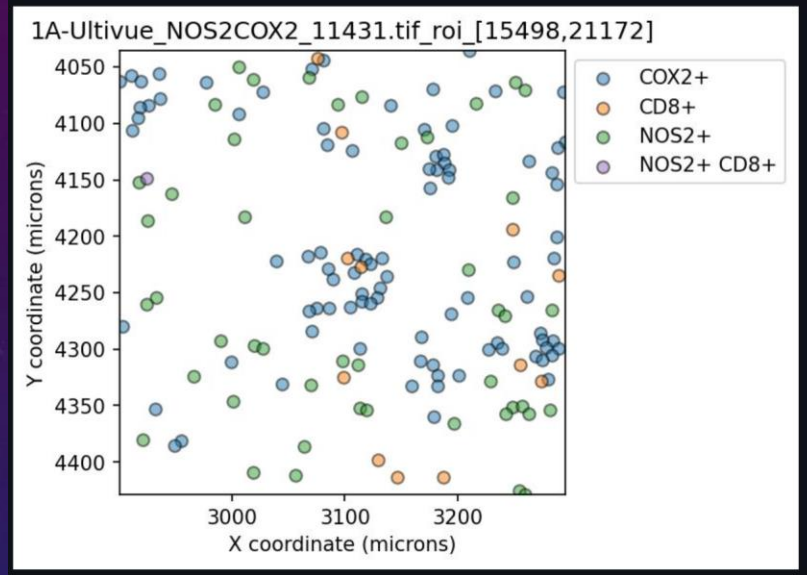


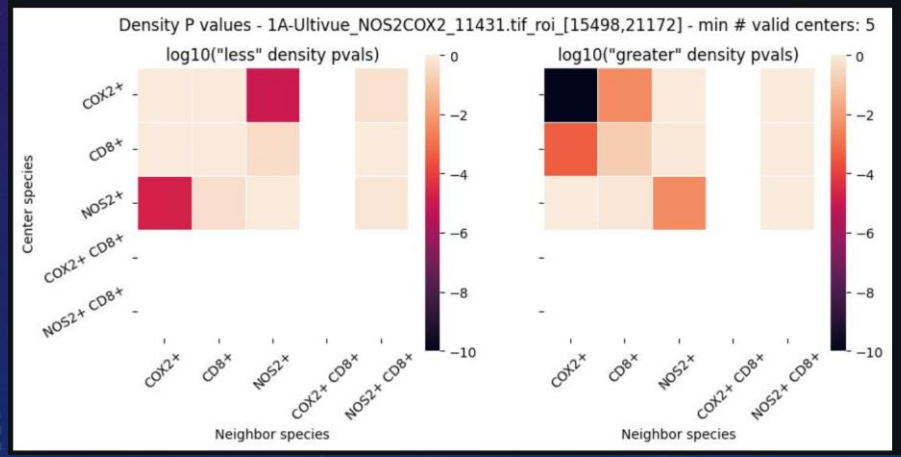
OTHER ANALYSIS SETTINGS

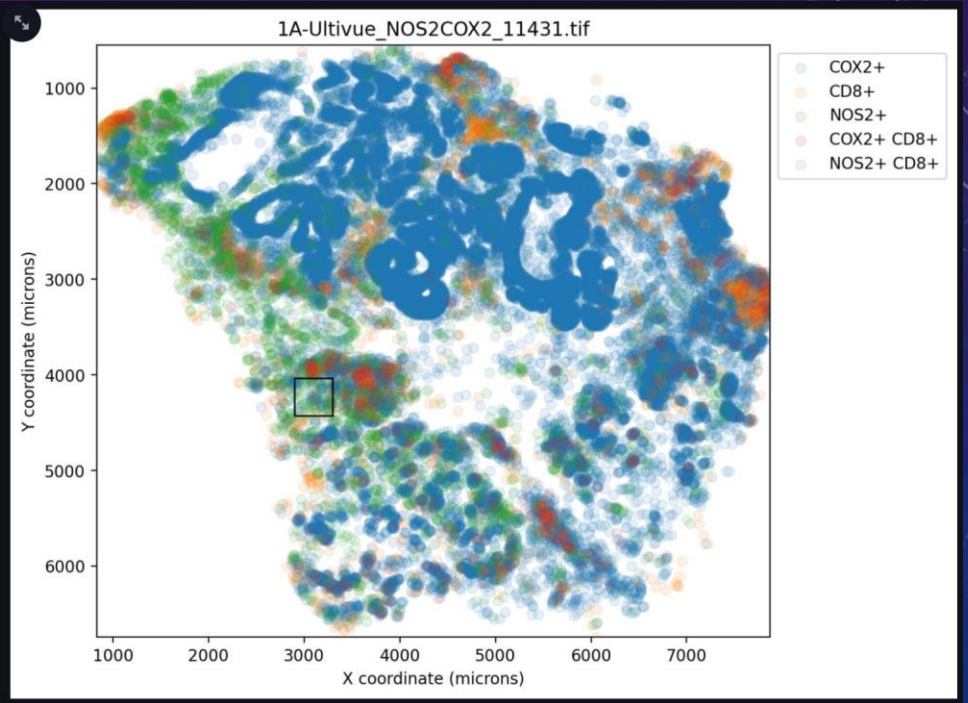
- We also implement a k-nearest neighbors, permutation-based method.
- Slides/images are generally heterogeneous.
- It's often best to patch them up into more homogeneous ROIs.
- These ROIs should be overlapping so cells discarded at the edge of one ROI are included in the analysis of an adjacent ROI.
- For this to work, the ROI overlap should be twice the analysis radius.



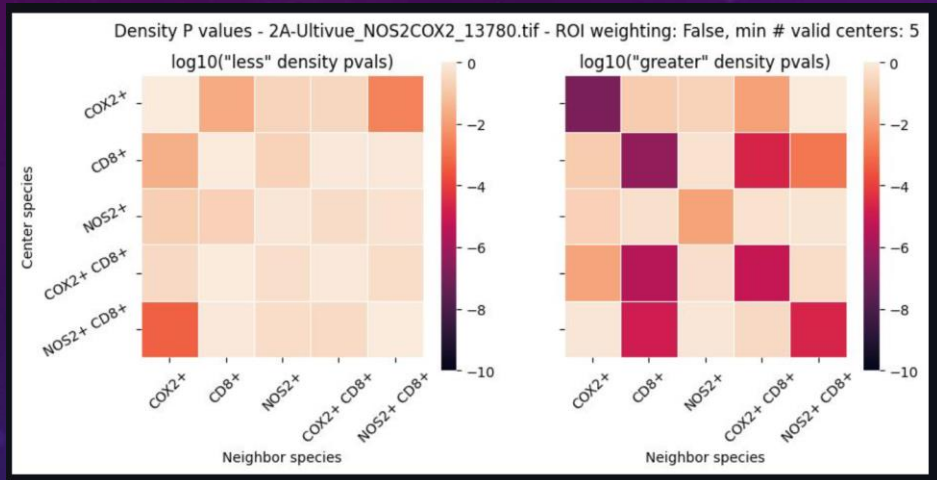
SAMPLE ROI-BASED OUTPUT

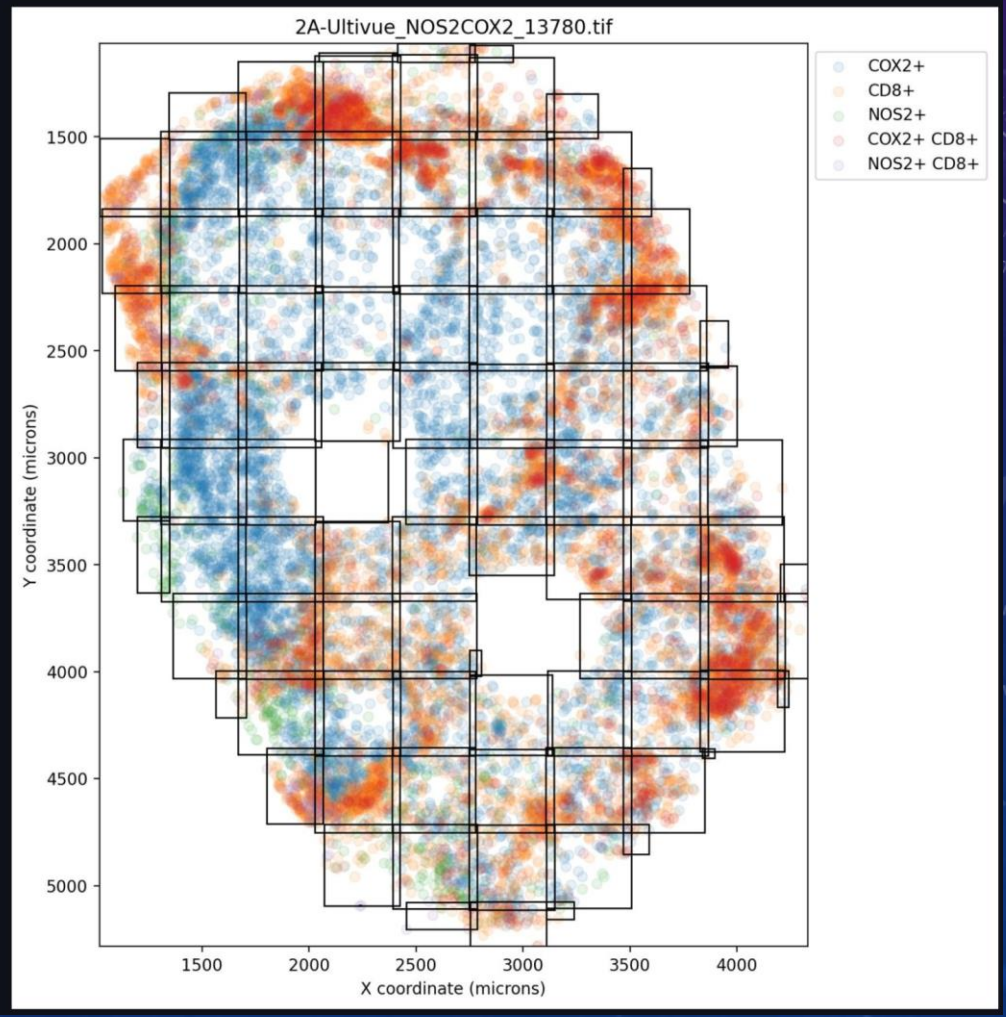




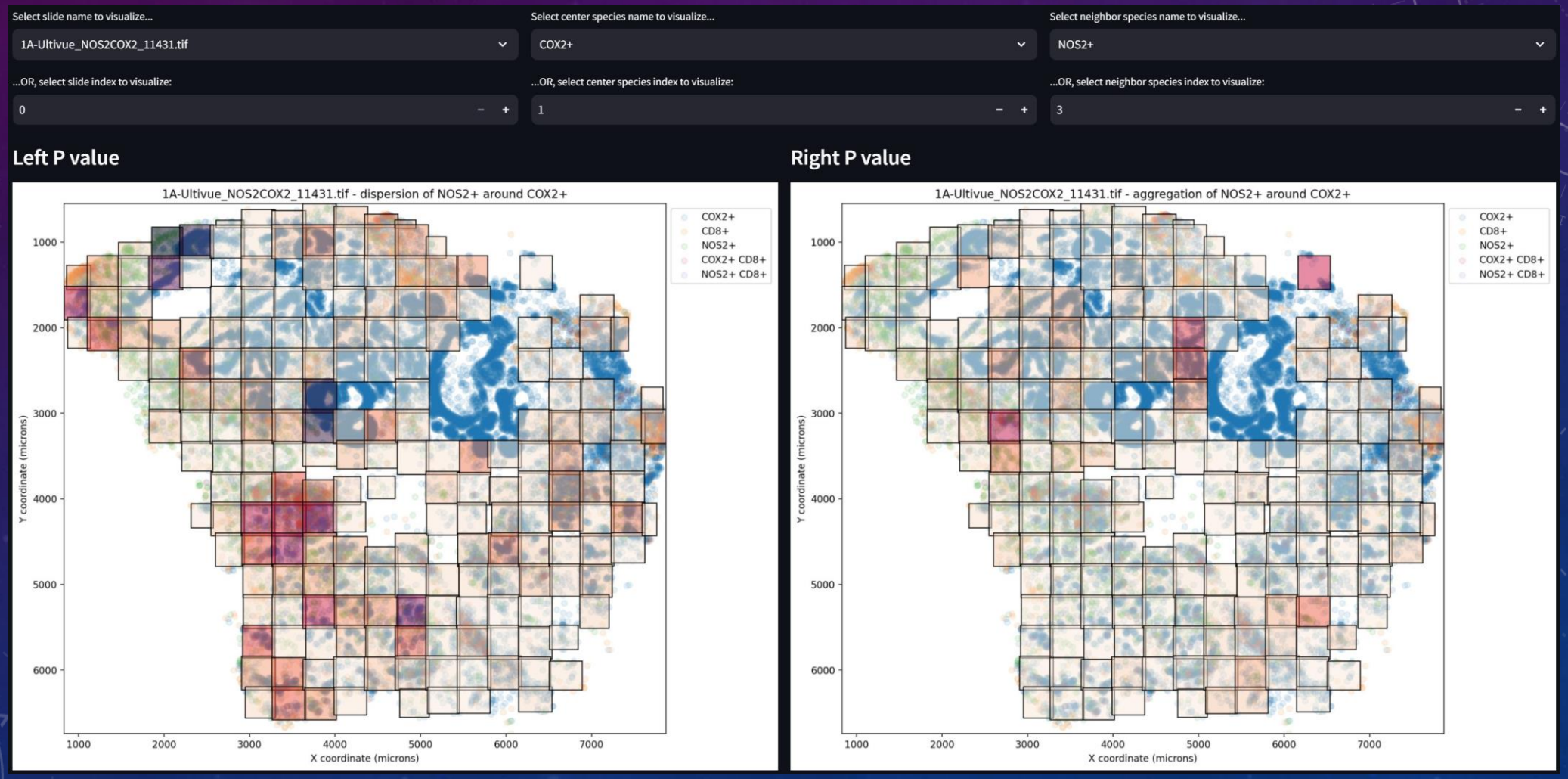


SAMPLE SLIDE-BASED OUTPUT



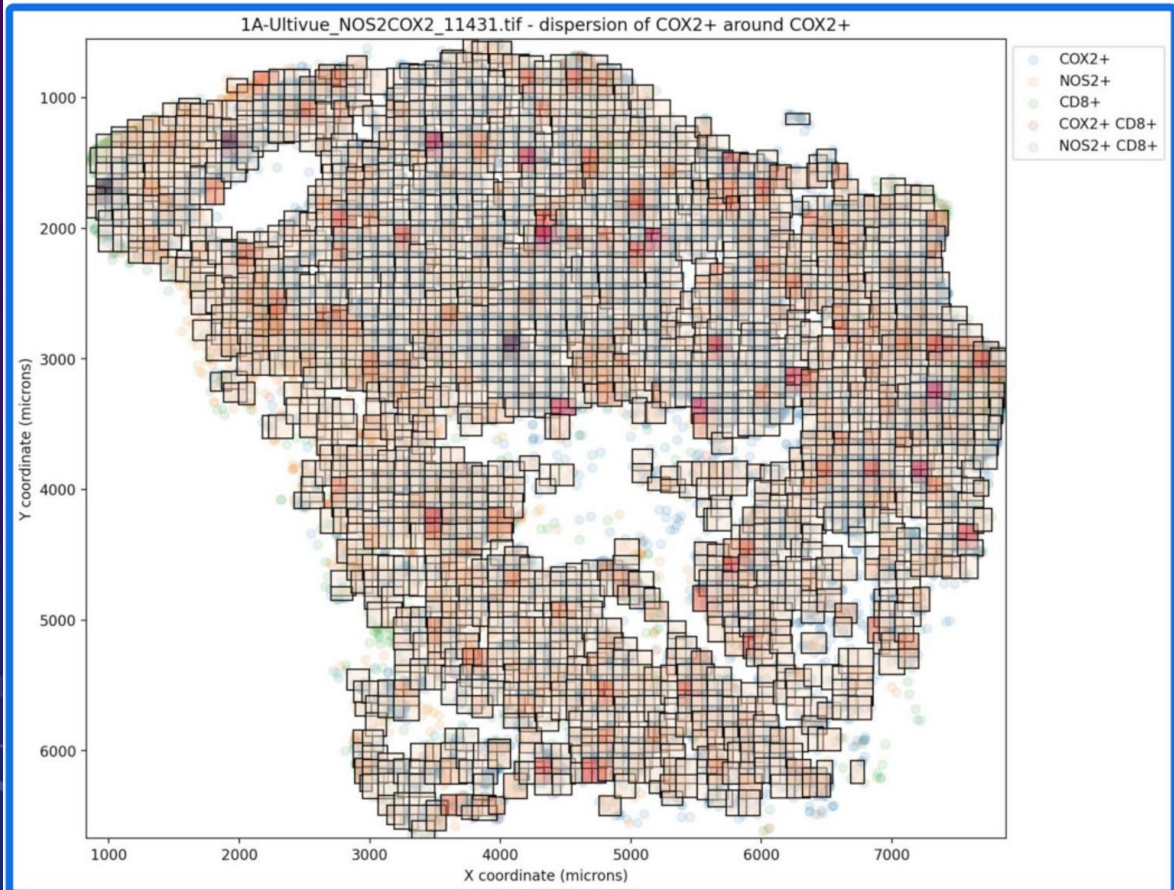


SAMPLE P VALUES OVERLAID ON SLIDES

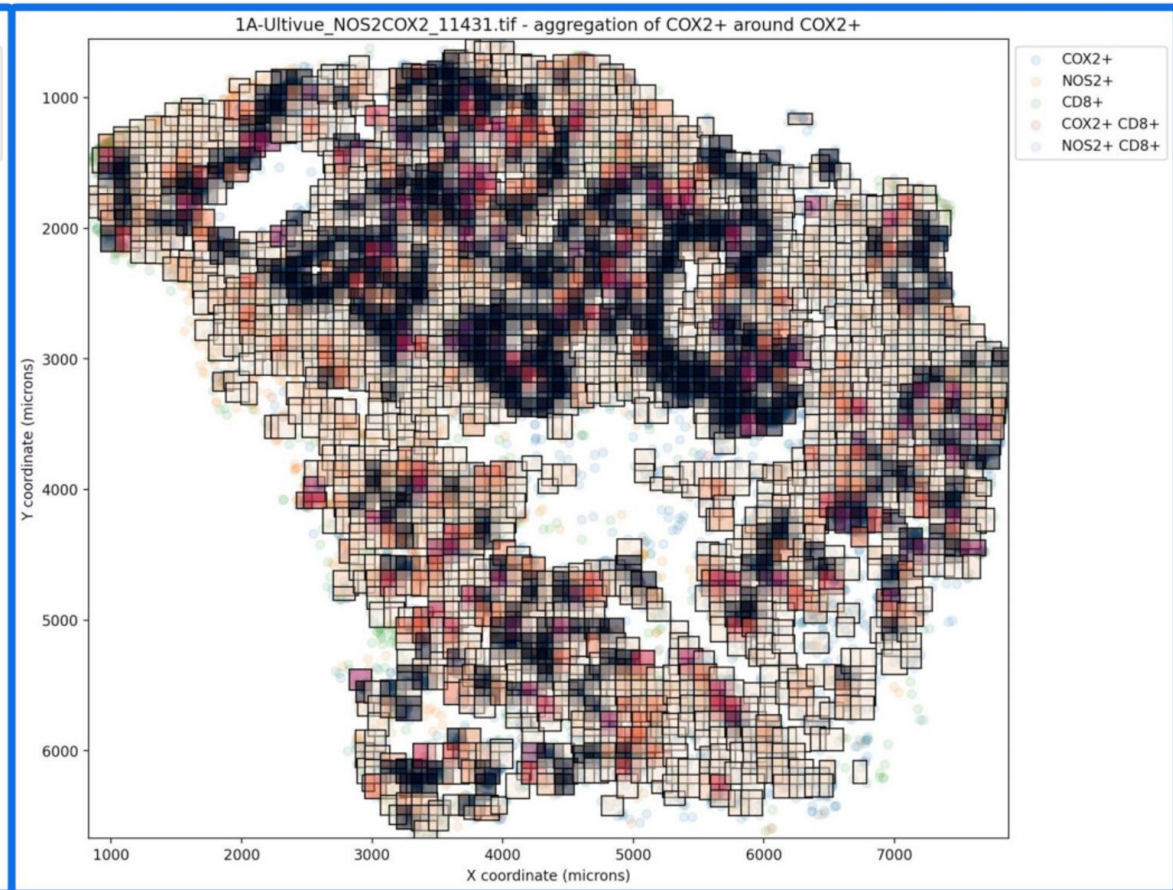


SAMPLE P VALUES OVERLAID ON SLIDES (SMALLER ROIS)

Left *P* value



Right *P* value



NEXT STEPS

- Homework:
  - Review the recording and script for this session of this worksho
  - Follow along step-by-step in the training MAWA deployment.
- Plan to attend the final session:
  - Wed 11/6, 11-12 PM
    - Neighborhood analysis using spatial UMAP.
    - Dante Smith, Ph.D.