Integration and multiplex of single cell RNA-seq

Neo Christopher Chung

Lecture 12, 1000-719bMSB

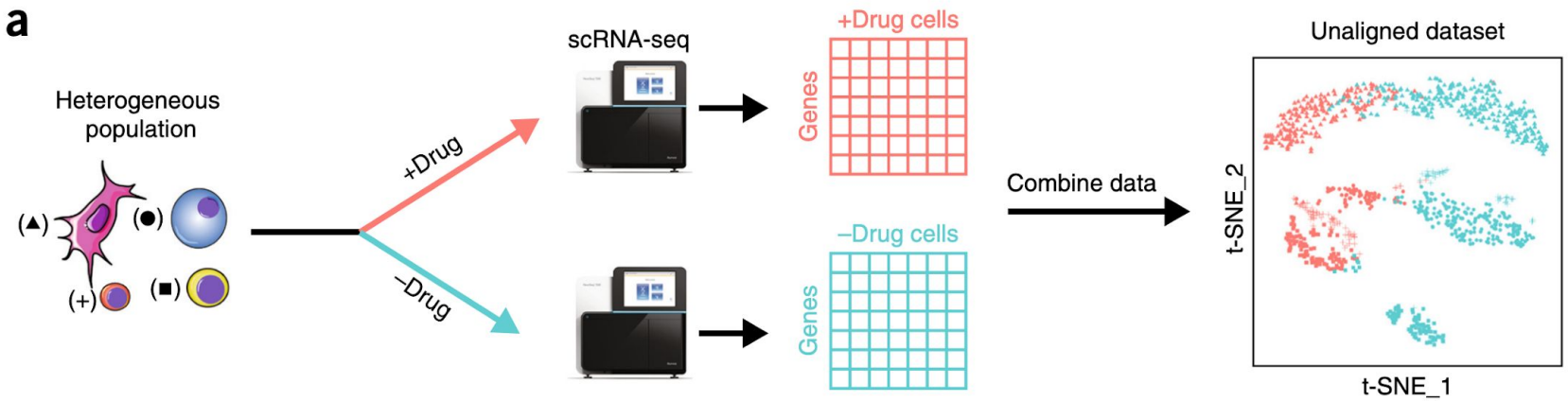
Why integrate single-cell RNA-seq Data?

1. How can disparate single-cell datasets be harmonized into a single reference?
2. Different technologies/platforms available for sequencing
3. Different environmental or experimental conditions
4. How can its reference data improve the analysis of new experiments?
5. Potential for the information present in one experiment to inform the interpretation of another

→ Bulter et al. (2018), Haghverdi et al., (2018), Stuart et al. (2019)

Challenges in integrating data from a case-control study

1. Cells are treated with two different agents (+Drug vs. -Drug)
2. These cells are sequenced en masse, where cell identities are unknown
3. Data are combined (naively)
4. Cells are clustered both by cell types and drug treatments
5. How to control for cell types that are not labeled at single cell levels?



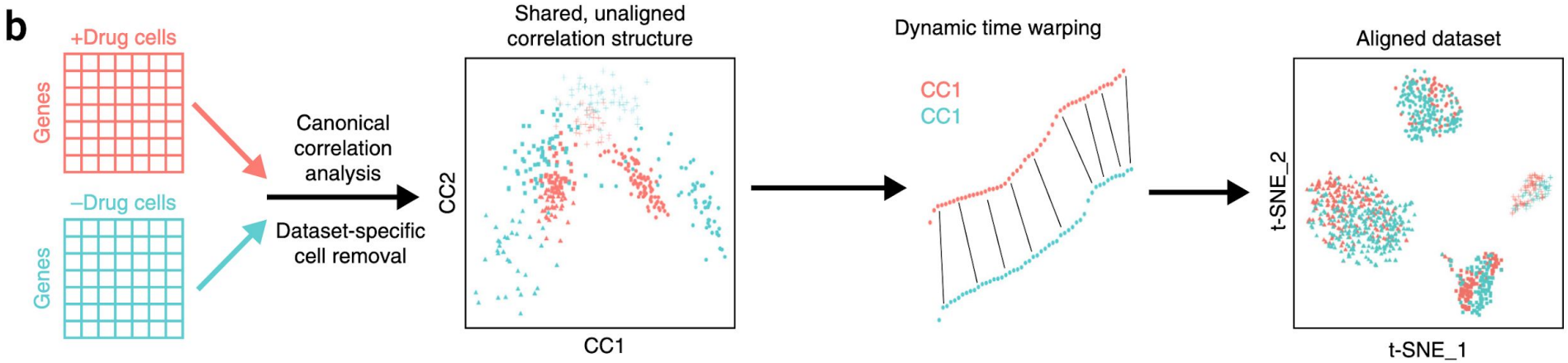
Four cell types are represented by different symbols, while drug treatment is encoded by color.

Improving interpretability of scRNA-seq

1. Integration using computational approaches
    a. canonical correlation analysis (CCA)
    b. dynamic time warping (DTW)
    c. mutual nearest neighbors (MNNs)
2. Multiplexing using experimental and computational approaches
    a. Demuxlet - using natural genetic variations
    b. Cell hashing - creating artificial genetic variations

Computational Approach with CCA (Bulter et al. 2018)

1. Identify the shared gene correlation structure, conserved between the data sets using canonical correlation analysis (CCA)
2. Select/remove individual cells that cannot be well described by this shared structure
3. Align the data sets into a conserved low-dimensional space, using dynamic time warping (DTW) algorithms



Four cell types are clustered separately. Within each cell type, treatment effects are shown.

Canonical correlation analysis

1. CCA aims to find linear combinations of features across data sets that are maximally correlated, identifying shared correlation structures across data sets
2. CCA has been used for multimodal genomic analysis from bulk samples, for example, identifying relationships between gene expression and DNA copy number measurements based on the same set of samples
3. CCA treats the data sets as multiple measurements of a gene–gene covariance structure, and search for patterns that are common to the data sets.

(Classical) canonical correlation analysis

Let $X_{g,c}$ be a gene expression matrix of genes $g_1, g_2, ..., g_n$ by cells $c_1, c_2, ..., c_m$

$Y_{g,d}$ be a gene expression matrix of the same genes $g_1, g_2, ..., g_n$ by cells $d_1, d_2, ..., d_p$.

$$\max_{u,v} u^T X^T Yv, \text{ subject to } u^T X^T Xu \leq 1 \text{ and } v^T Y^T Yv \leq 1$$

Essentially, we are analyzing the cross-covariance matrix given by:

$$\mathrm{K_{XY}} = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}[X_1])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(Y_2 - \mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_1 - \mathrm{E}[X_1])(Y_n - \mathrm{E}[Y_n])] \\ \mathrm{E}[(X_2 - \mathrm{E}[X_2])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(Y_2 - \mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_2 - \mathrm{E}[X_2])(Y_n - \mathrm{E}[Y_n])] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_m - \mathrm{E}[X_m])(Y_1 - \mathrm{E}[Y_1])] & \mathrm{E}[(X_m - \mathrm{E}[X_m])(Y_2 - \mathrm{E}[Y_2])] & \cdots & \mathrm{E}[(X_m - \mathrm{E}[X_m])(Y_n - \mathrm{E}[Y_n])] \end{bmatrix}$$

However, the number of genes of interest that are shared between the two data sets is often much smaller than the total number of cells that were measured ($n << m + p$)

Diagonalized canonical correlation analysis

However, the number of genes of interest that are shared between the two data sets is often much smaller than the total number of cells that were measured ($n << m + p$).

Consequently, the vectors u and v that are returned from CCA will not be unique.

Treat the covariance matrix within each data set as diagonal.

Substitute the identity matrix for X^TX and Y^TY to arrive at

$$\max_{u,v} u^T X^T Yv, \text{ subject to } \|u\|_2^2 \leq 1 \text{ and } \|v\|_2^2 \leq 1$$

Note that the constraints are changed slightly, which leads to more efficient computation.

Solving diagonalized canonical correlation analysis

Note to standardize columns (samples) of X and Y to have a mean of 0 and variance of 1.

Then, apply singular value decomposition (SVD) on a cross-covariance matrix of X and Y:

$$K_{XY} = X^T Y$$

$$K_{XY} = \Gamma \Lambda \Delta^T$$

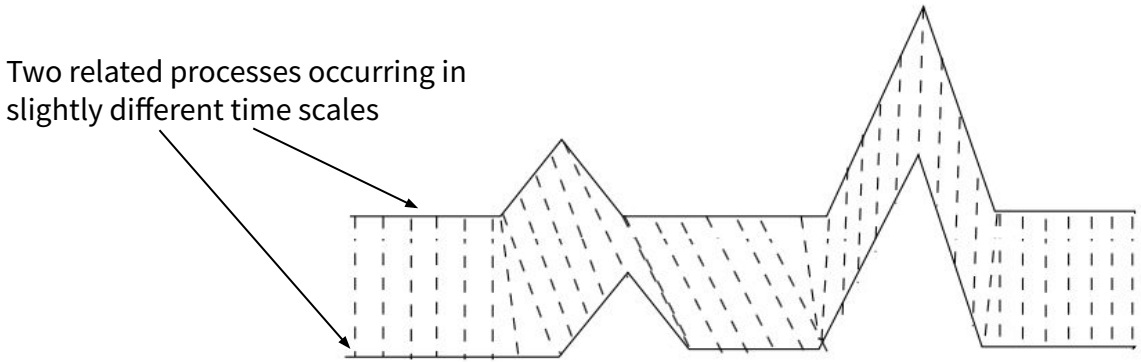
where $\Gamma = (\gamma_1, \ldots, \gamma_k)$; $\Delta = (\delta_1, \ldots, \delta_k)$; $\Lambda = (\lambda_1, \ldots, \lambda_k)$

Canonical correlation vectors u and v are simply the left and right singular vectors from the SVD:

$$u_i = \gamma_i \quad \text{and} \quad v_i = \lambda_i$$

Dynamic Time Warping

1. Represent each basis vector as a metagene, defined as a weighted expression average of the top genes whose expression exhibits robust correlation with the basis vector
2. Linearly transform the metagenes to match their 95% reference range, correcting for global differences in feature scale.
3. Determine a mapping between the metagenes using dynamic time warping, which locally compresses or stretches the vectors

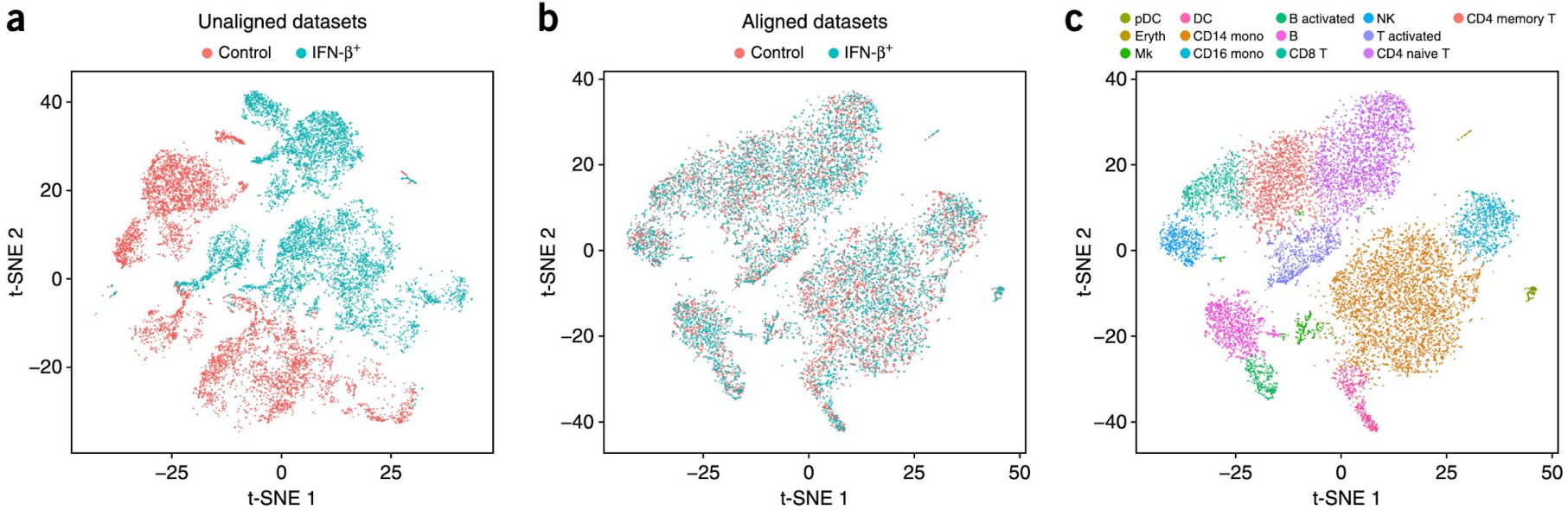


Application on stimulated and resting PBMCs

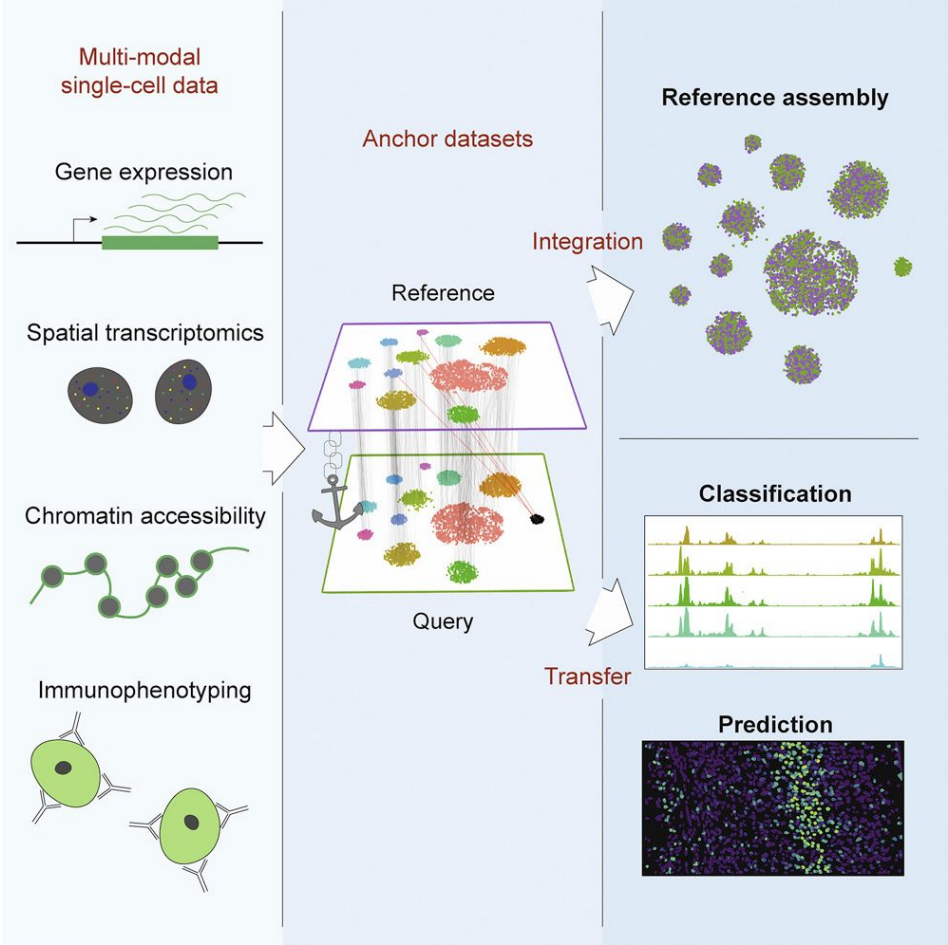
14,039 human PBMCs from eight patients into two groups:

one **stimulated** with interferon-beta (IFN-β)

another culture-matched **control**

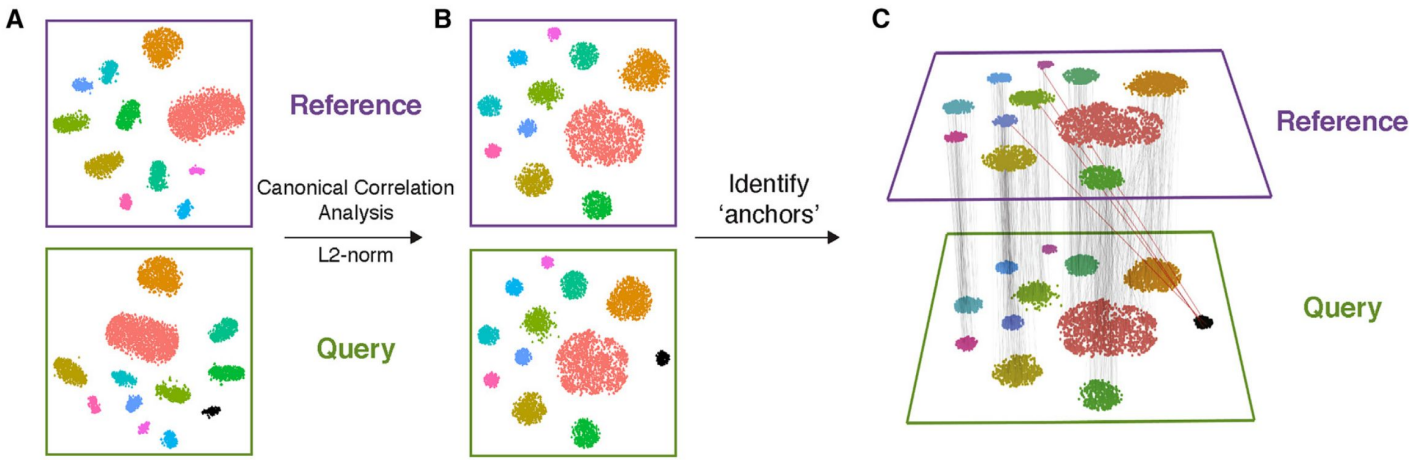


Improved pipeline from Stuart et al. (2019)



CCA + MNN approach

1. Get normalized CCA
2. Get mutual nearest neighbors (MNNs) as anchors
3. Use MNNs to correct for batch effects

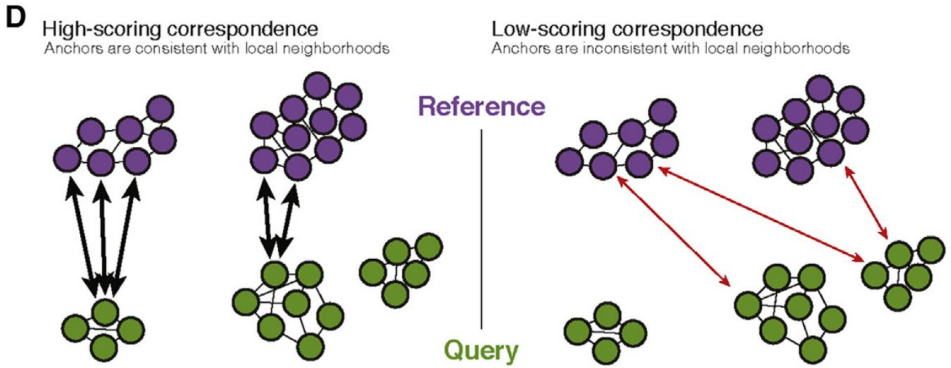


CCA + MNN approach

1. Jointly reduce the dimensionality of both datasets **Y** using diagonalized CCA
2. Apply L2-normalization to the canonical correlation vectors
   E.g., Divide the vectors by a square root of a sum of squares
3. Search for mutual nearest neighbors (MNNs) in this shared low-dimensional representation (resulting cell pairs are called **anchors**)
4. Each anchor pair was assigned a score based on the shared overlap of mutual neighborhoods for the two cells in a pair
5. Calculate the anchor weight matrix **W,** the strength of association between each query cell c, and each anchor i. Based on the distance between the query cell and the anchor, and the previously computed anchor score
6. Batch correction by $\mathbf{Y} - \mathbf{BW}^T$, where B is the difference between the two expression vectors for every pair of anchor cells

Mutual nearest neighbors from Haghverdi et al. (2018)

1. Identify the K-nearest neighbors (KNNs) for each cell within its paired dataset, based on the L2-normalized CCV
2. Identify mutual nearest neighbors (MNN), which are pairs of cells, with one from each dataset, that are contained within each other's neighborhoods
3. Each anchor pair was assigned a score based on the shared overlap of mutual neighborhoods for the two cells in a pair



Anchor scoring

1. Goal: anchors identified in low-dimensional space are supported by original high-dimensional measurements.
    a. Examine the nearest neighbors of each anchor query cell in the reference dataset.
    b. If the anchor reference cell is found within the **first k.filter (200) neighbors**, then we retain this anchor.
    c. Otherwise, we remove this anchor from further analyses.
2. Goal: Minimize the influence of incorrectly identified anchors
    a. For each reference anchor cell, we determine its k.score (30) nearest within-dataset neighbors and its k.score nearest neighbors in the query dataset.
    b. Combine to form an overall neighborhood graph
    c. Compute the shared neighbor overlap between the anchor and query cells, and assign this value as the **anchor score**
    d. Use the 0.01 and 0.90 quantiles to rescale anchor scores to a range of 0 to 1

Experimental approach to help integration

1. Difficulty with figuring out unknown cell identities
2. Computational weakness in post-hoc or meta-analysis
3. Better to prepare the eventual need for integration by experimental approach

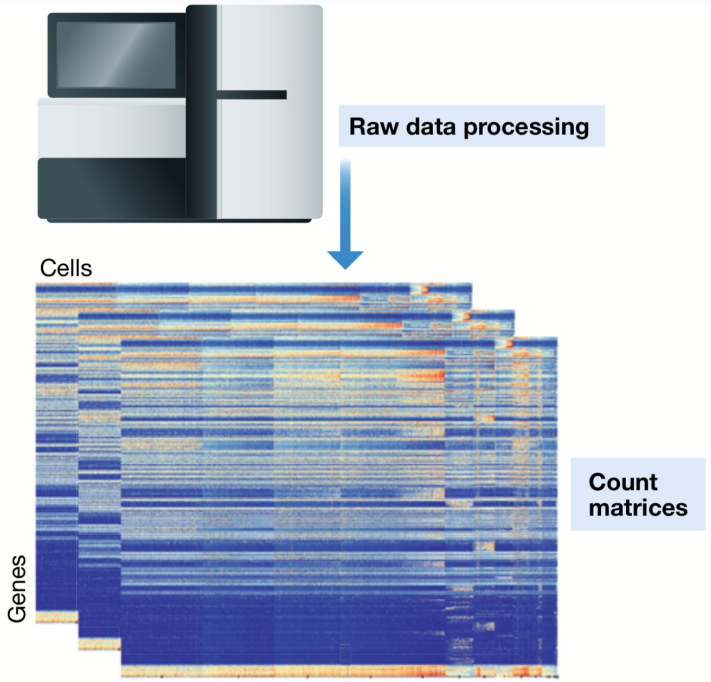
Sample Multiplexing

Reliable identification of doublets and multiplets

Experimental approaches to sequence multiple conditions (e.g., case vs. control)

Easily and reliably distinguish the origins of multiple samples in a single run/lane

Note that this doesn't completely solve the problem of identifying (known and unknown) cell types



Harnessing Genetic Variation Kang HM et al (2017)

Useful when multiple samples are from different genetic backgrounds

Retaining the origins of those single cells through a statistical model

It requires pooled samples to originate from previously genotyped individuals

Demuxlet

demuxlet enables the pooling of samples with distinct genotypes together into a single scRNA-seq experiment.

The sample-specific genetic polymorphisms serve as a fingerprint for the sample of origin and therefore can be used to assign each cell to an individual after sequencing

Maximum likelihood to determine the most likely donor for each cell using a mixture model

Cell Hashing from Stoeckius et al. (2018)

Extending demuxlet by combining with **oligonucleotide-tagged antibodies**

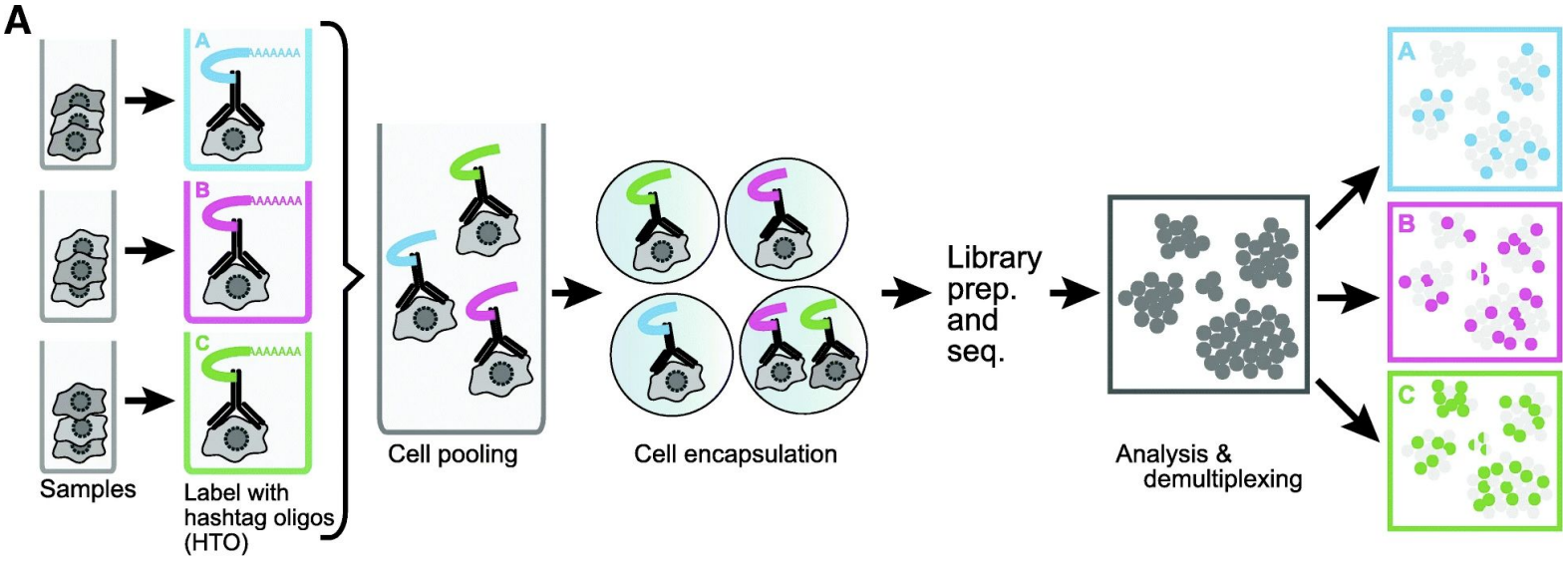
Widely used in flow cytometry and mass-cytometry (CyTOF)

Oligonucleotide-tagged antibodies are used to convert the detection of cell surface proteins into a sequenceable readout alongside scRNA-seq

Based on the concept of hash functions in computer science to index datasets with specific features, a set of oligo-derived hashtags equally define a “lookup table” to assign each multiplexed cell to its original sample

Cell Hashing

1. Cells from different samples are incubated with DNA-barcoded antibodies recognizing ubiquitous cell surface proteins.
2. Distinct barcodes (referred to as hashtag-oligos, HTO) on the antibodies allow pooling of multiple samples into one scRNA-seq experiment.
3. After sequencing, cells can be assigned to their sample of origin based on HTO levels



Benchmark with PBMCs from 8 donors

1. Peripheral blood mononuclear cells (PBMCs) from eight separate human donors (referred toas donors A through H)
2. Chose a set of monoclonal antibodies directed against ubiquitously and highly expressed immune surface markers (CD45,CD98, CD44, and CD11a)
3. Pooled all cells together in equal proportion, alongside an equal number of unstained HEK293T cells (and 3%mouse NIH-3T3 cells) as **negative controls**
4. The HTOs contain a unique 12-bp barcode that can be sequenced alongside the cellular transcriptome

HTO classification algorithm for Cell Hashing

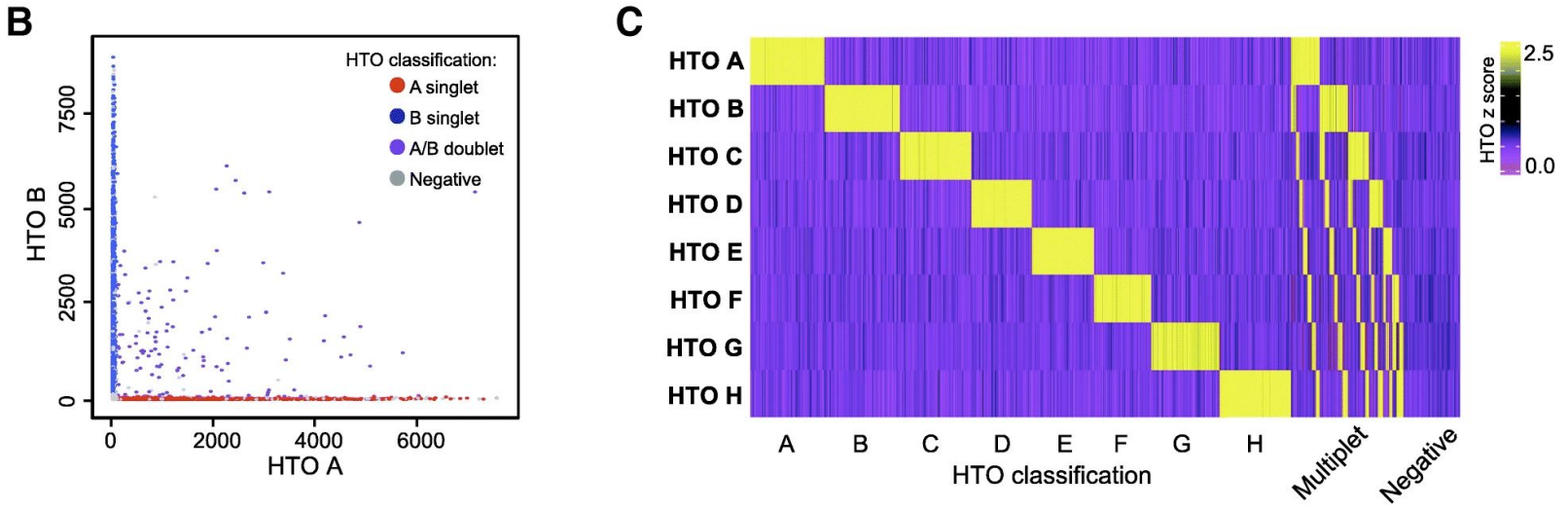
1. HTO raw counts were normalized using centered log ratio (CLR) transformation:

$$x_i' = \log \frac{x_i}{\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}}$$

2. Perform k-medoids clustering of all HTO reads with K = 9.
   a. 8 clusters enriched for expression of a particular HTO, while 9th cluster enriched for cells with low expression of all HTOs
3. For each of the eight HTOs, model the“background”:
   a. identify the k-medoids cluster with the highest average HTO expression and excluded these cells
   b. exclude the highest 0.5% values as potential outliers
   c. fit a negative binomial distribution to the remaining HTO values
   d. calculated the q= 0.99 quantile of the fitted distribution and thresholded each cell in the dataset based on this HTO-specific value

HTO classification algorithm for Cell Hashing

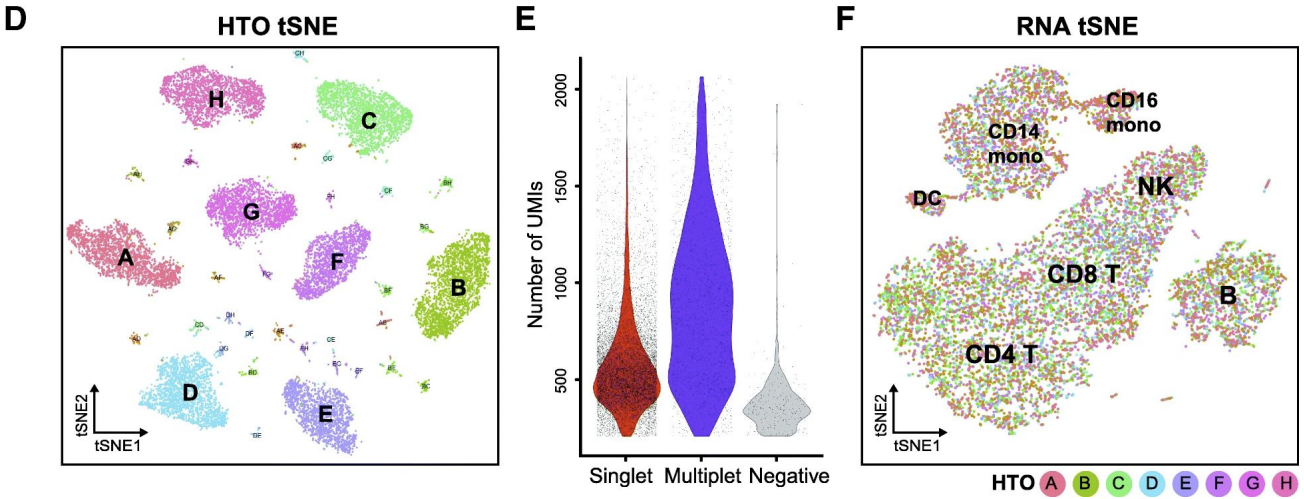
4. Barcodes that were positive for only one HTO were classified as singlets.
5. Barcodes that were positive for >1 HTOs were classified as multiplets.
6. Barcodes that were negative for all eight HTOs were classified as “negative.”



t-SNE projections from HTO and RNA

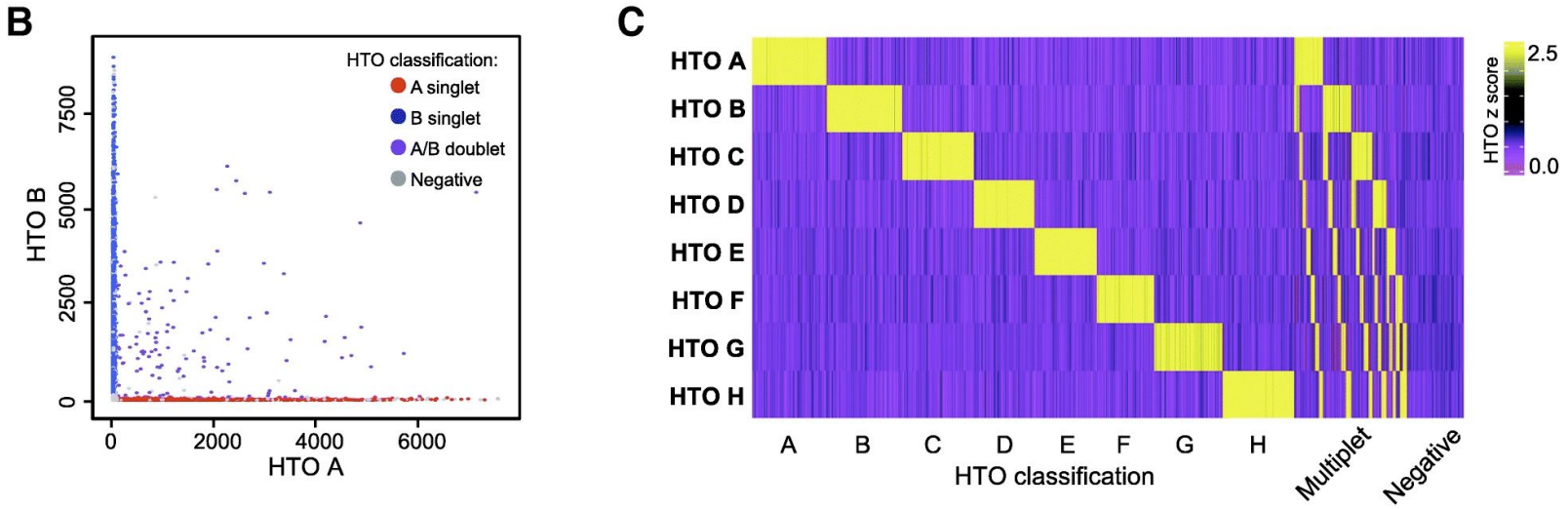
For HTO t-SNE (D), use euclidean distances calculated for tSNE
For RNA t-SNE (F), use the top 10 PCs of the 1000 most highly variable genes

In both, their computationally determined HTO classifications are shown.



Analyzing DNA-barcoded antibodies

Cell Hashing algorithm is a procedure.
How can we assign statistical significance and posterior probabilities to these?



Jackstraw P-values for cluster memberships in HTO data