Ecole thématique sinceIITE 2019

Statistical models and analysis

2nd part

Antonio Rausell, Ph.D.

Roscoff, February 6th 2019

imagine
INSTITUT DES MALADIES GÉNÉTIQUES

The bioinformatics pipeline: main “modular” components

Yesterday

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

Today

8 - Batch Effect correction and data integration

9 - Single-cell matching across datasets

antonio.rausell@institutimagine.org

The bioinformatics pipeline: main “modular” components

Yesterday

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

7- **A note on statistical robustness**

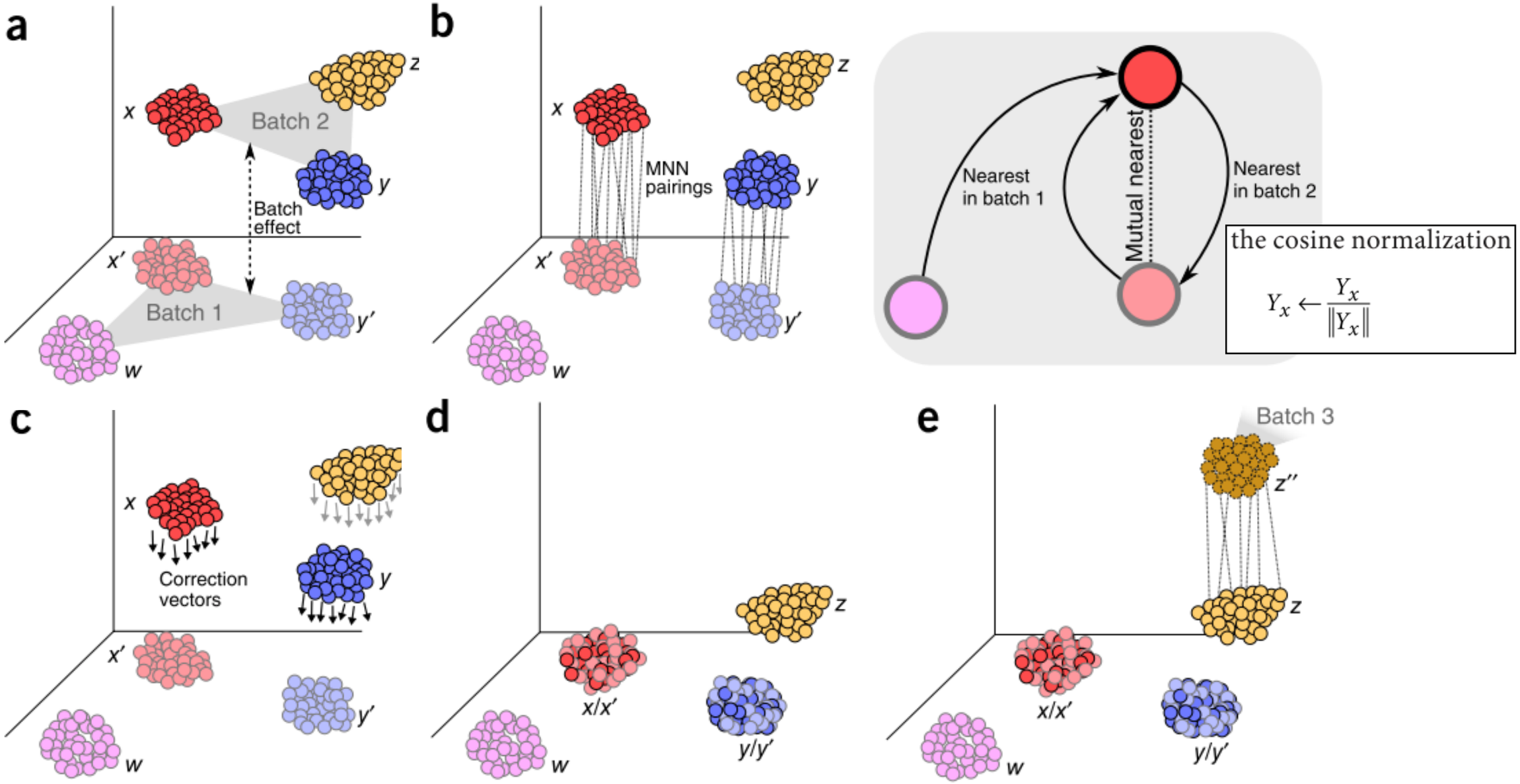
Today

8 - Batch Effect correction and data integration

9 - Single-cell matching across datasets

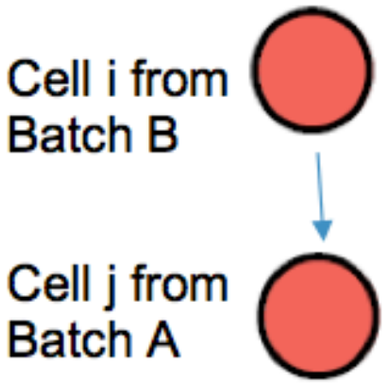
antonio.rausell@institutimagine.org

Mutual Nearest Neighbors (MNN): (I) Overview



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

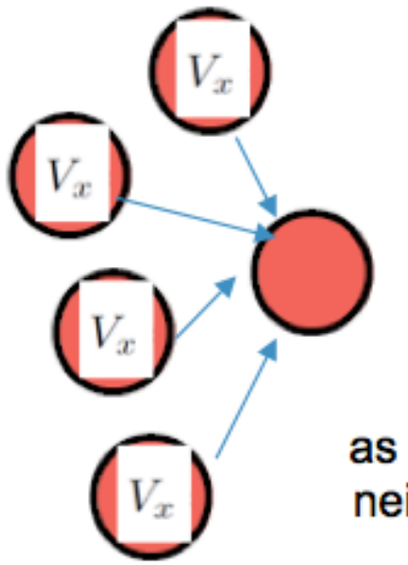
Mutual Nearest Neighbors (MNN): (II) Correction vectors



1) For each MNN pair, a pair-specific batch-correction vector is computed as the vector difference between the expression profiles of the paired cells.

$$V_x = \begin{pmatrix} gene1_a - gene1_b \\ gene2_a - gene2_b \\ gene3_a - gene3_b \\ \cdots \\ geneN_a - geneN_b \end{pmatrix}$$

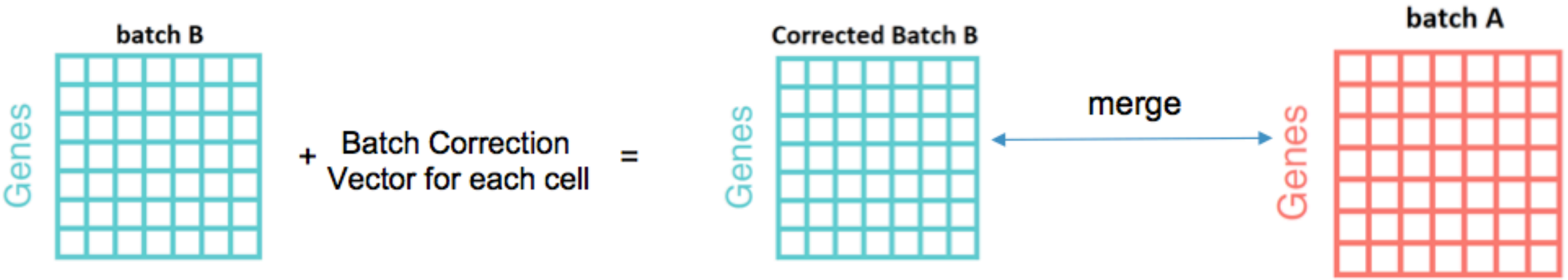
2) A cell-specific batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with a Gaussian kernel.



Gaussian Kernel Smoothing

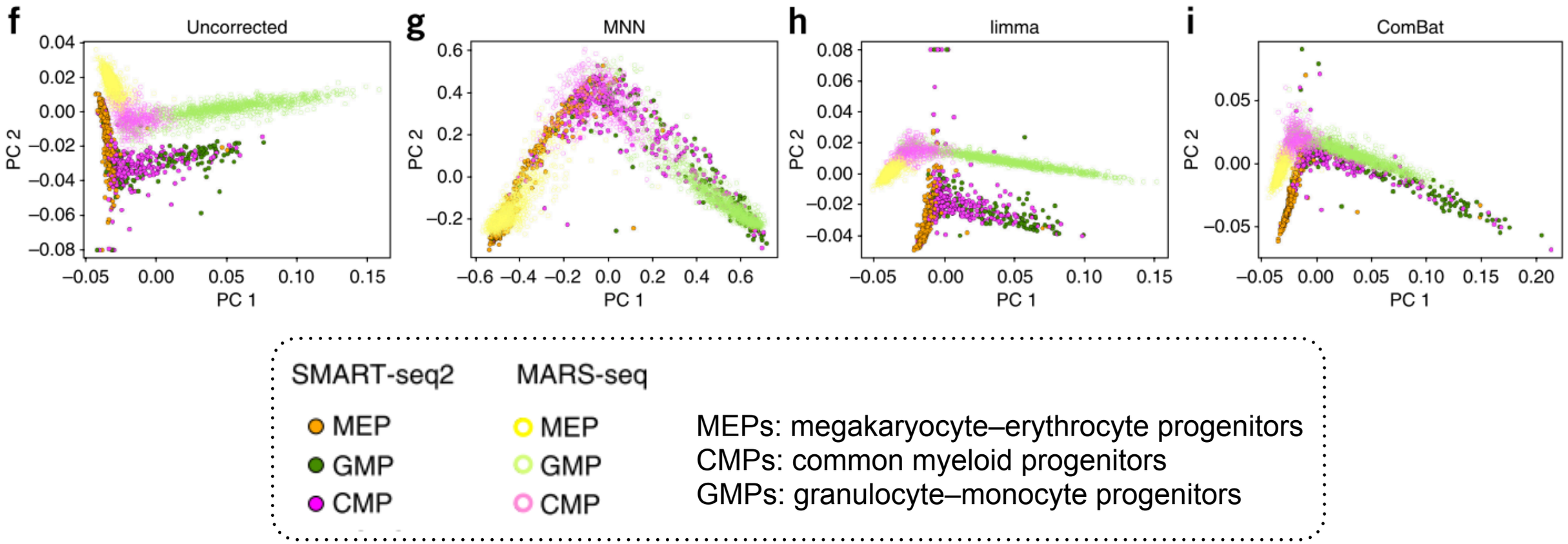
Real valued function
$f : R^p \rightarrow R$
as the weighted average of neighboring observed data

Batch Correction vector for each cell



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

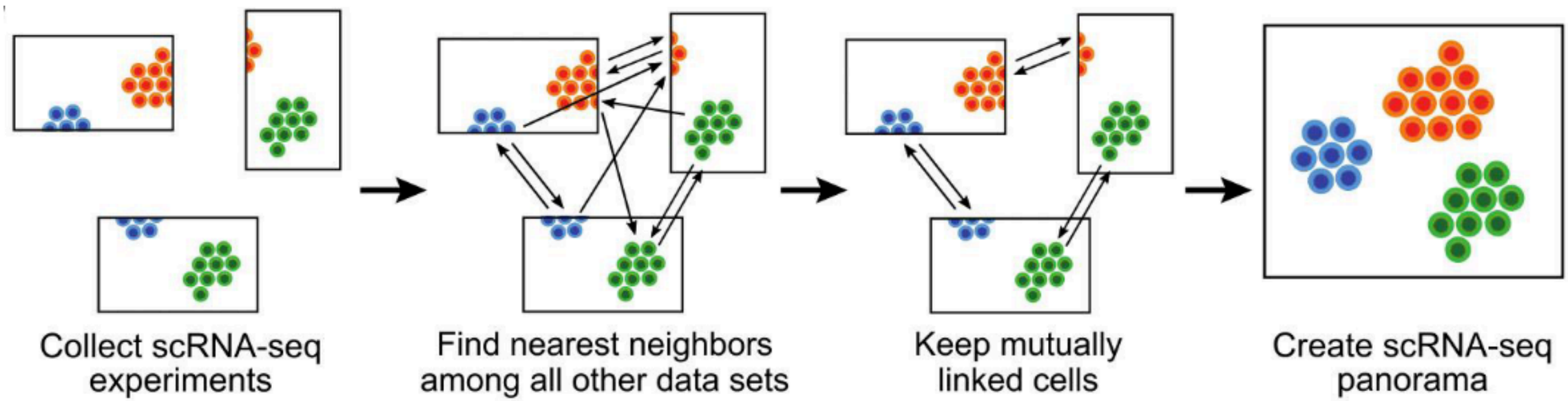
Mutual Nearest Neighbors (MNN): (III) Example



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.
Haghverdi et al. Nature Biotechnology 2018 doi:10.1038/nbt.4091

Scanorama: MNN-like method with focus on scalability

Panoramic stitching of heterogeneous single-cell transcriptomic data. Hie et al. BioRxiv 2018.
https://www.biorxiv.org/content/10.1101/371179v1

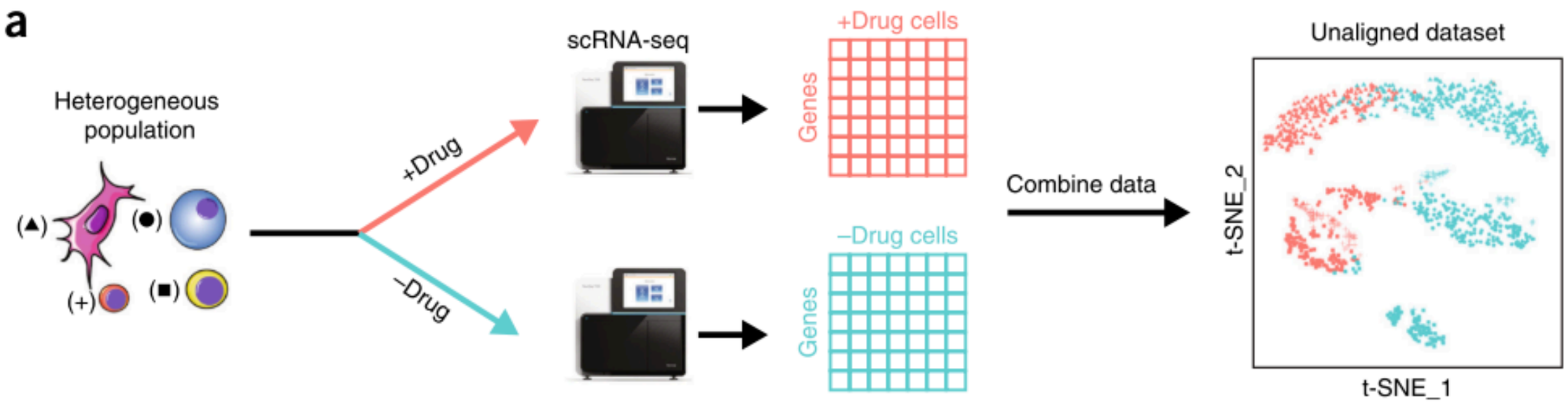


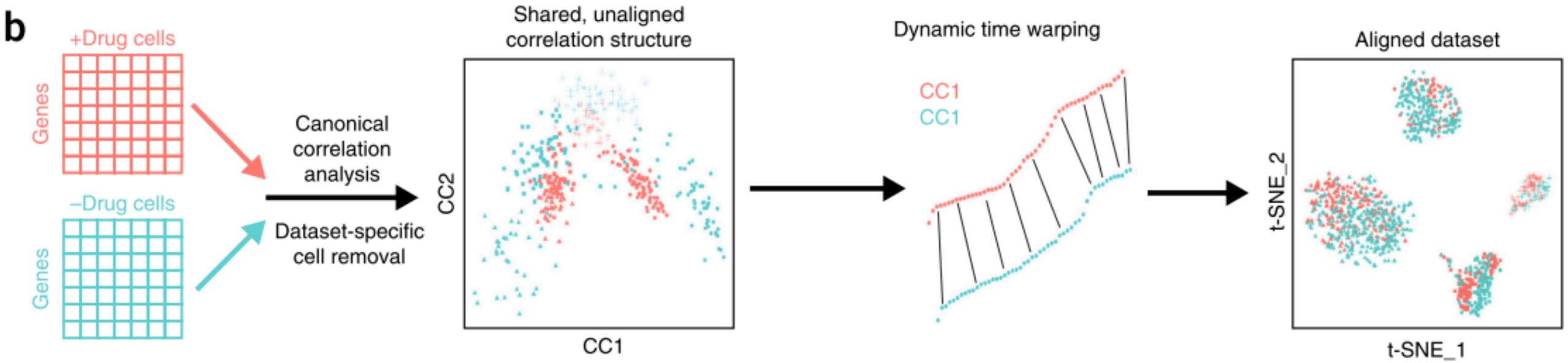
Focus on computational efficiency:

[...] Instead of performing the nearest neighbor search in the high-dimensional gene space, we compress the gene expression profiles of each cell into a low-dimensional embedding using an efficient, randomized singular value decomposition (SVD) (10) of the cell-by-gene expression matrix. [...]

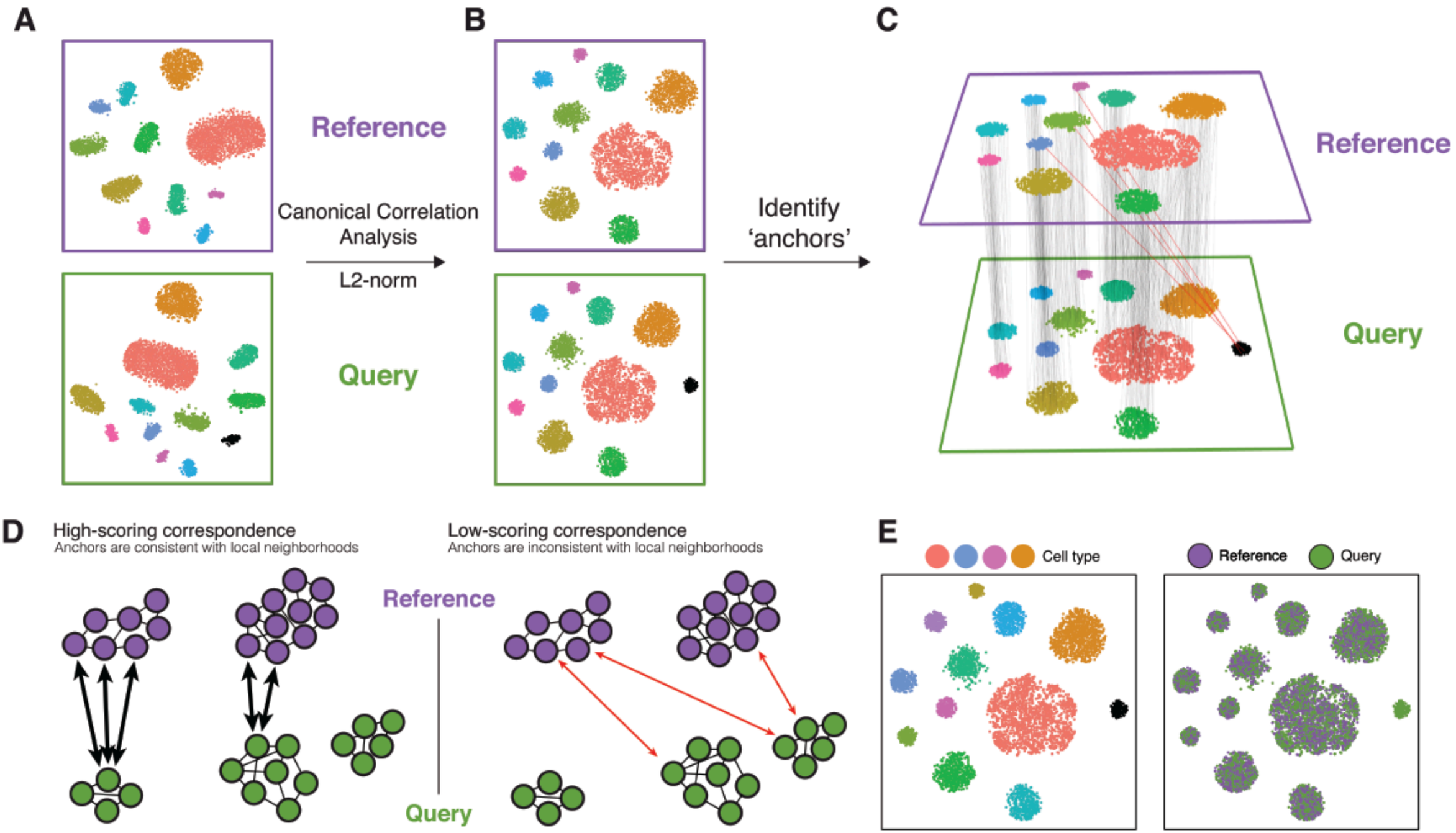
[...] Additionally, we use an approximate nearest neighbor search based on hyperplane locality sensitive hashing (11) and random projection trees (12) to greatly reduce the nearest neighbor (44) query time both asymptotically and in practice [...]

Seurat v2's integration: CCA + dynamic time warping



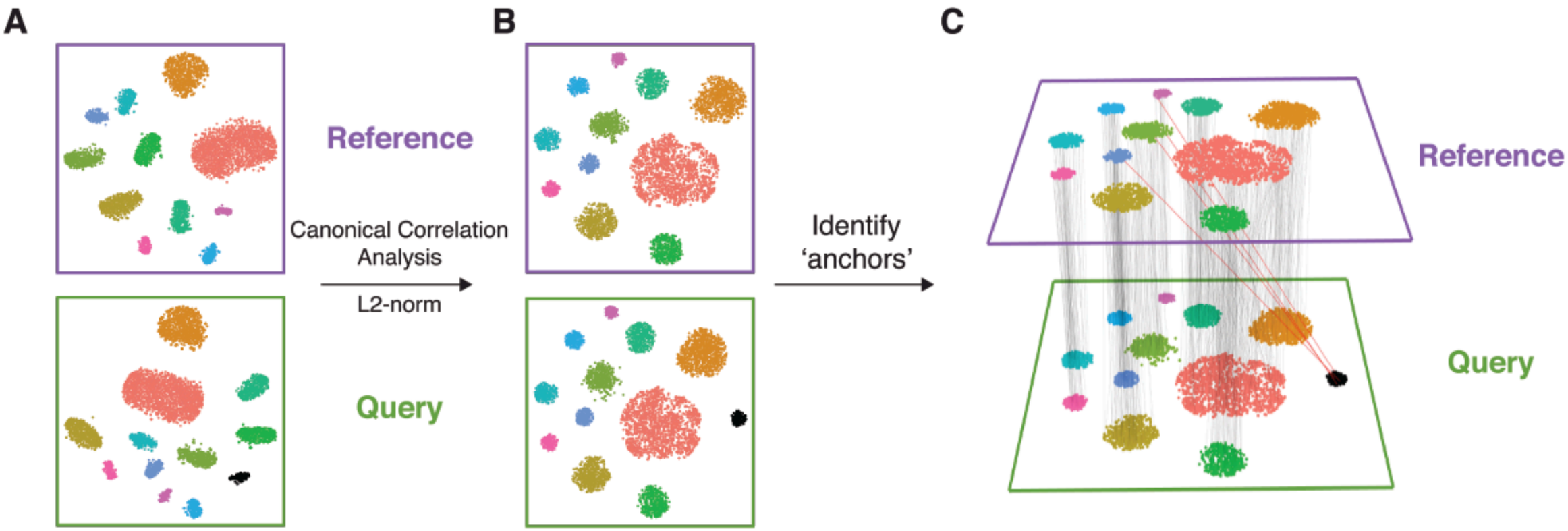


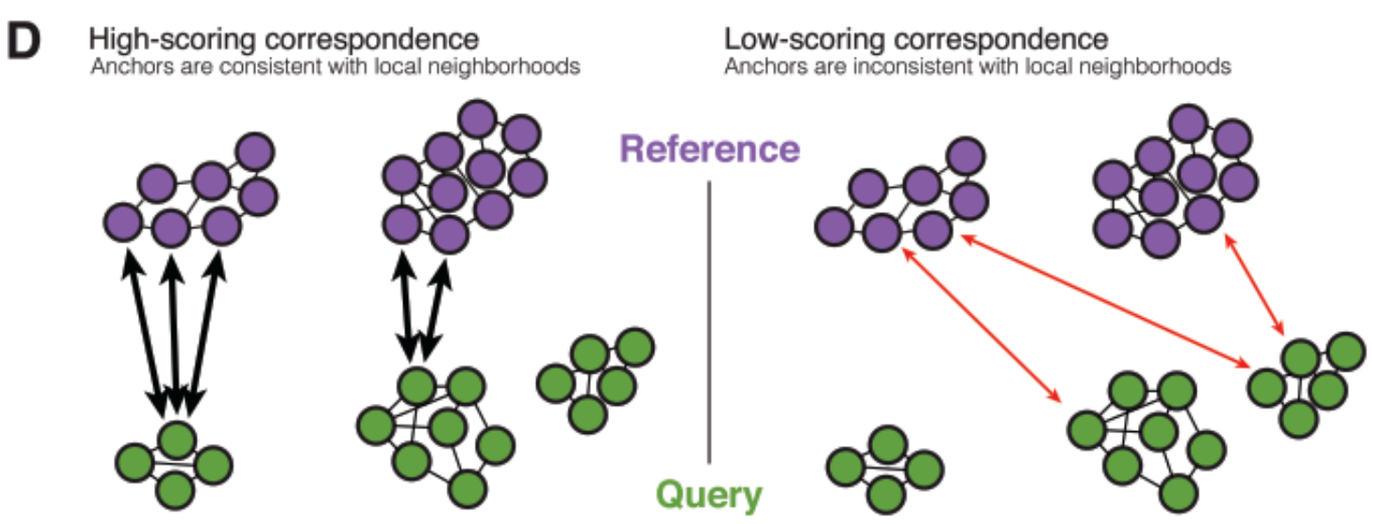
Seurat v3's integration: CCA + "anchors"



Comprehensive integration of single cell data Stuart et al. BioRxiv (2018)
https://www.biorxiv.org/content/10.1101/460147v1

Seurat v3's integration: CCA + "anchors"





1) calculate the matrix B, where each column represents the difference between the two expression vectors for every pair of anchor cells, a:

$$B = X[,a] - Y[,a]$$

2) calculate a transformation matrix, C, using the previously computed weights matrix and the integration matrix as:

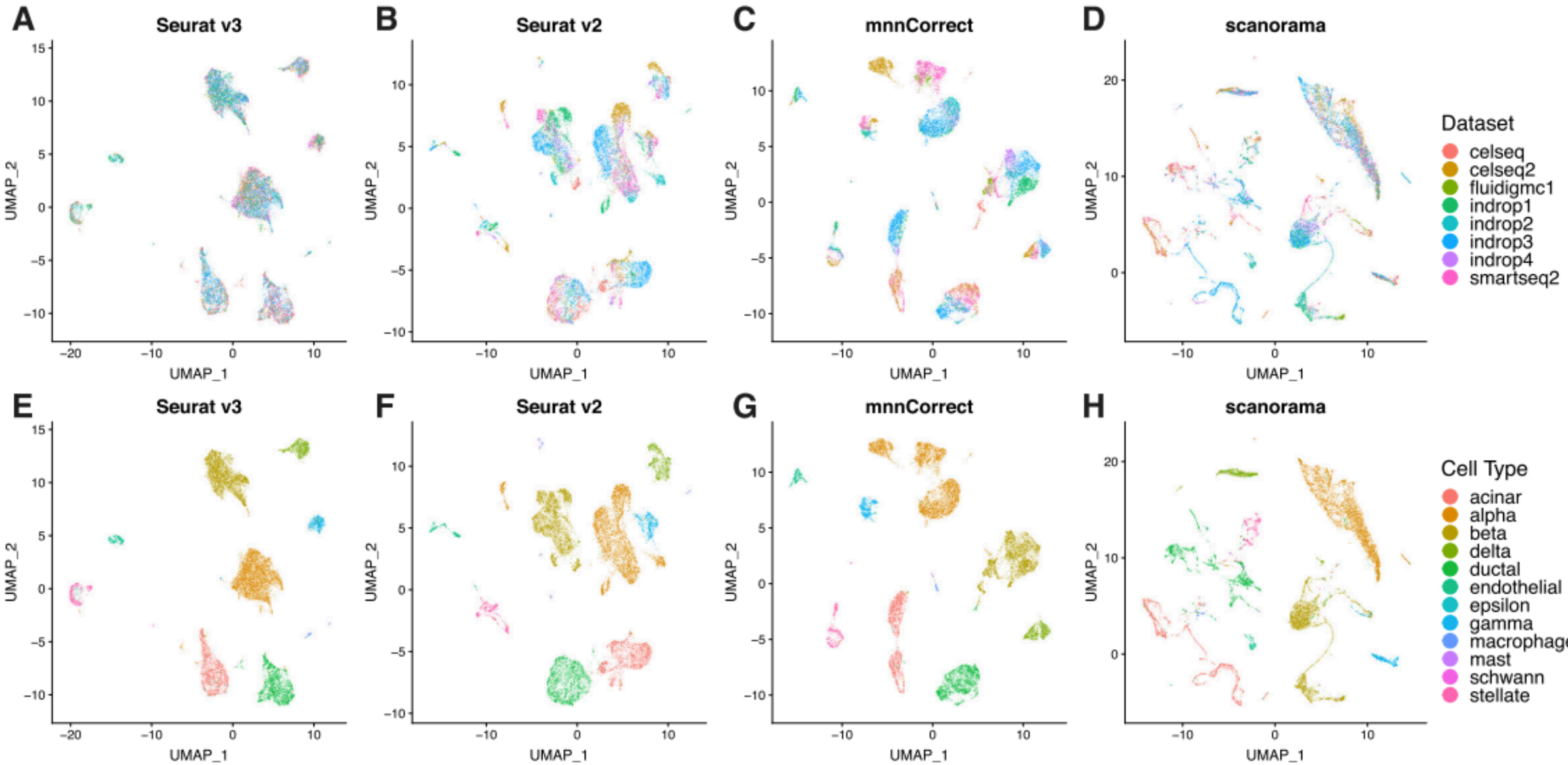
$$C = BW^T$$

3) subtract the transformation matrix, C, from the original expression matrix, Y, to produce the integrated expression matrix ˆY:

$$\hat{Y} = Y - C$$

Comprehensive integration of single cell data Stuart et al. BioRxiv (2018)
https://www.biorxiv.org/content/10.1101/460147v1

Seurat v3's integration: comparative performance



Comprehensive integration of single cell data Stuart et al. BioRxiv (2018)
https://www.biorxiv.org/content/10.1101/460147v1

Summary of integrative methods (Tian et al. BioRxiv 2018)

Supplementary Table 4. Summary of integrative methods used to combine data from different protocols and scNRA-seq studies. Methods can be classified into batch effect correction - where a batch-corrected data matrix is output, adjustment where the batch effect is accounted for in the model and dimension reduction methods where components or factors summarizing the batch-corrected data are output. Their hyperparameters are listed (*italic* indicates default parameters). HVG stands for Hyper-Variable Genes. SEG stands for Stably Expressed Genes.

Method	Correct	Adjust	Dim. reduction	# genes	Main parameters	Ref
MNN	✓			HVG	- *Number of nearest neighbors* - *Bandwidth of smoothing kernel*	[12]
MINT supervised			✓	all genes	- Number of components - If gene selection: Number of genes to select	[40]
ZINB-WaVe	✓	✓		HVG	- Number of factors	[37]
diagonal CCA		✓	✓	HVG	- Number of components - Reference dataset - *If multiCCA: number of iterations*	[43]
scMerge unsupervised	✓			all genes (+ SEG)	- Number of K-means clusters - *Number of factors* - *Ratio of pseudo replicates* - *Distance metric*	[28]
Scanorama	✓			HVG	- *Number of HVG* - *Number of nearest neighbors (NN)* - *Choice of approximate kNN* - *Gaussian kernel function parameter*	[14]

scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. Tian et al. BioRxiv 2018 https://www.biorxiv.org/content/10.1101/433102v2

“Benchmark” of integrative methods (Tian et al. BioRxiv 2018)

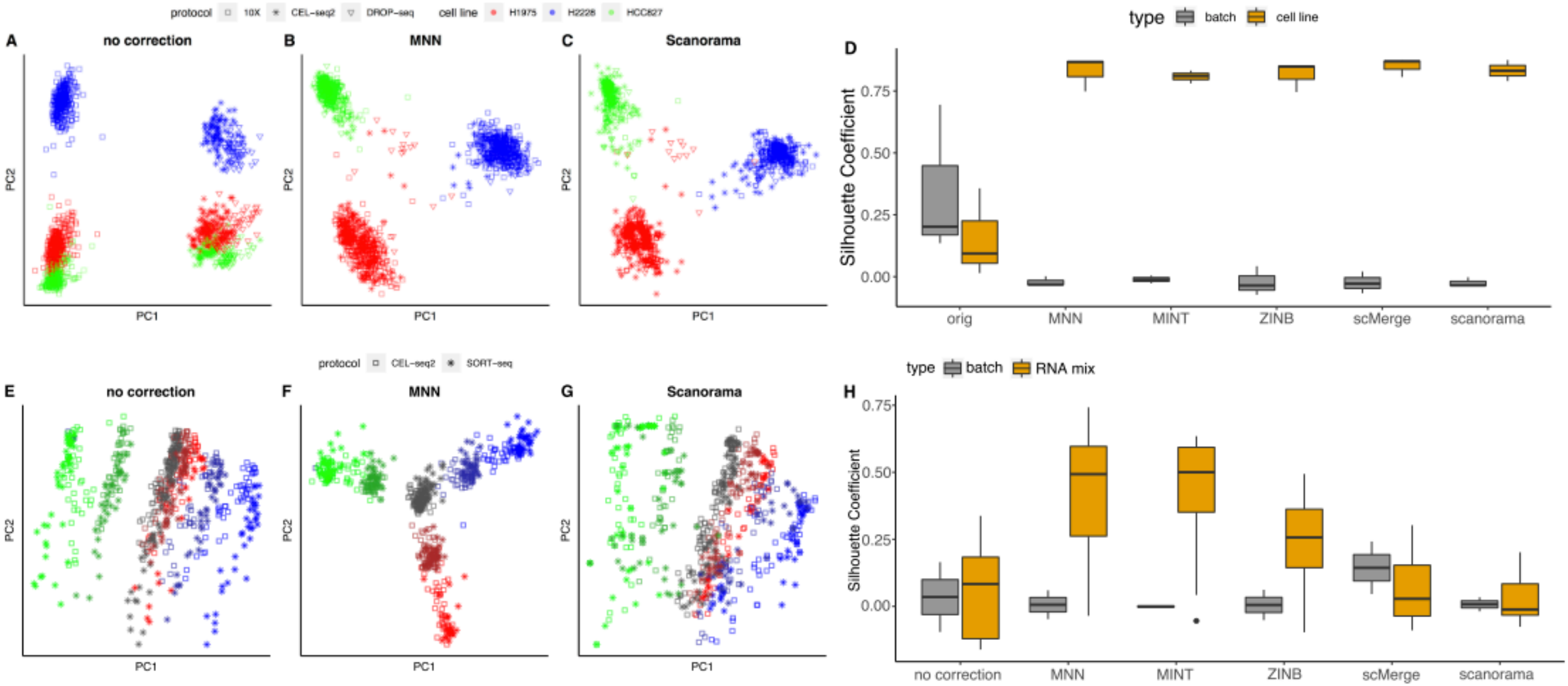
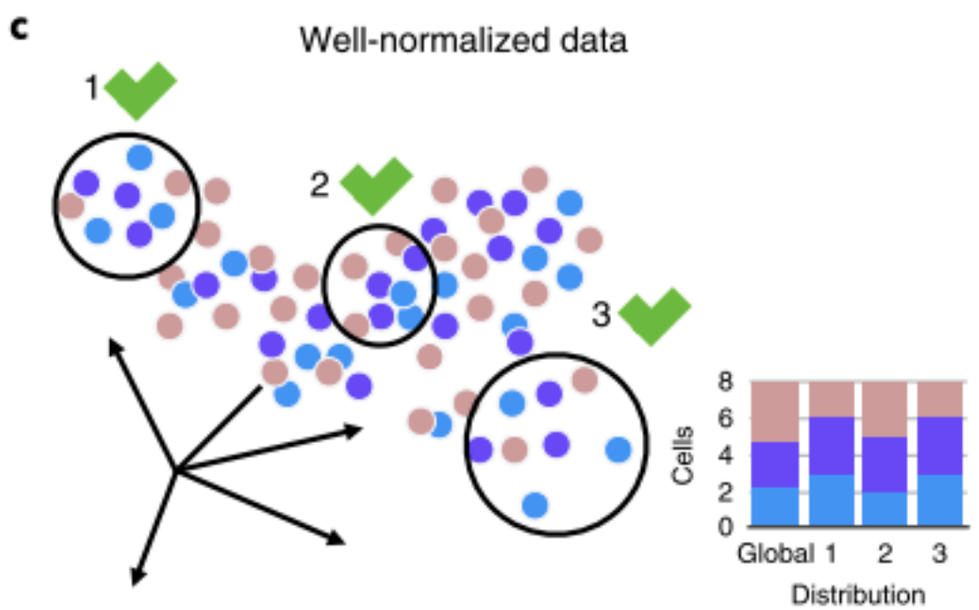
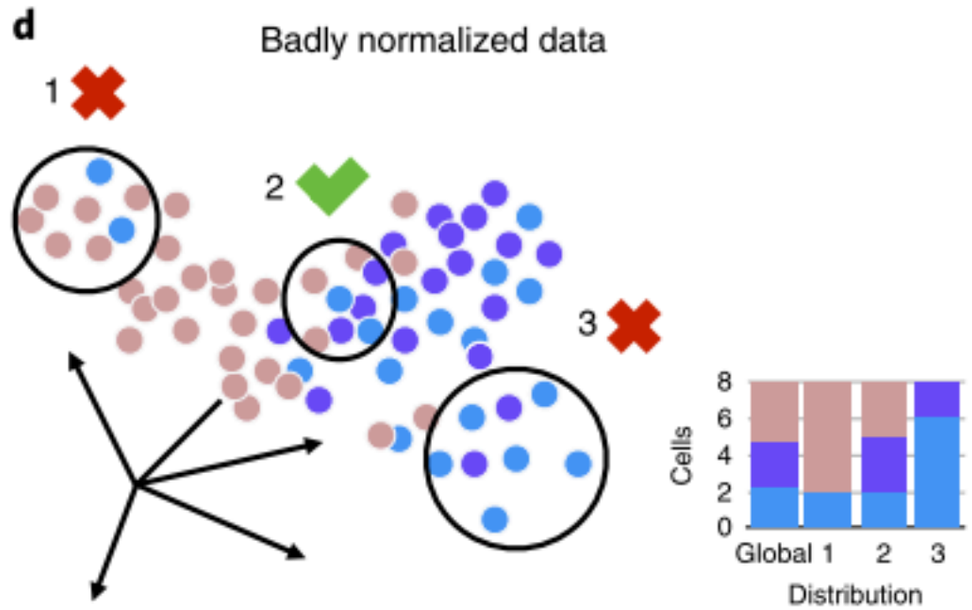


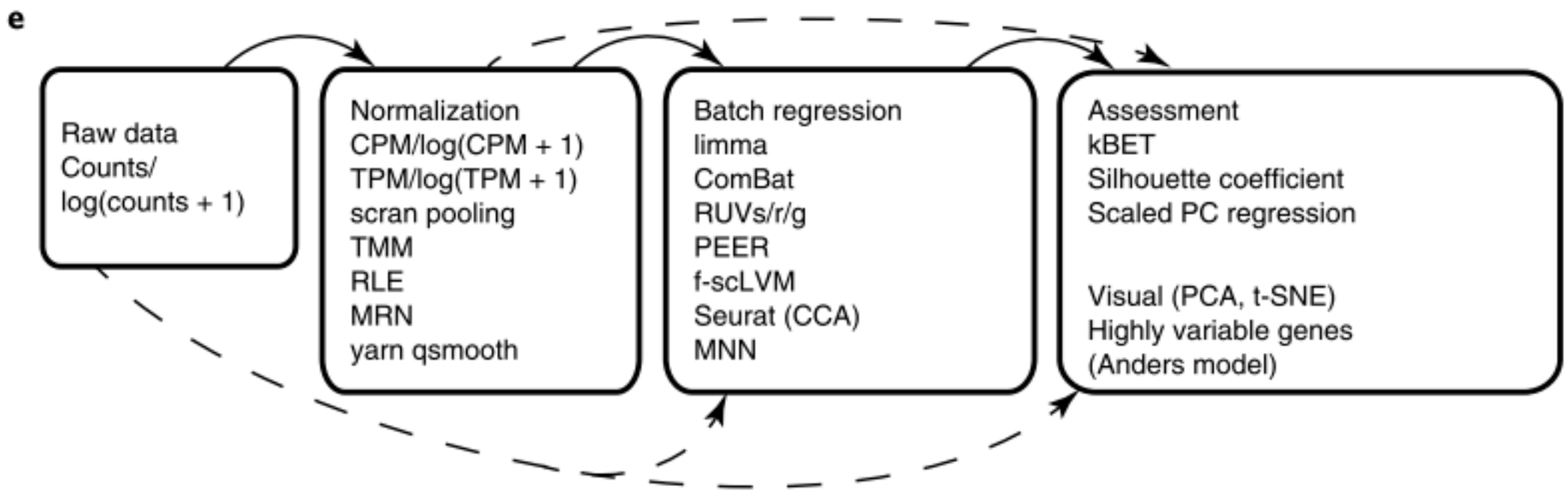
Figure 5. Comparisons of data integration methods for batch effect correction for the three single cell experiments and the RNA mixture experiments. (**A**,**E**) PCA sample plot when the

scRNA-seq mixology: towards better benchmarking of single cell RNA-seq protocols and analysis methods. Tian et al. BioRxiv 2018 https://www.biorxiv.org/content/10.1101/433102v2

Never mind... assess the quality of the integration in your own data with kBET







A test metric for assessing single-cell RNA-seq batch correction
Büttner et al. Nature Methods 2019. http://www.nature.com/articles/s41592-018-0254-1

The bioinformatics pipeline: main “modular” components

Yesterday

1- Feature selection

2- Dimensionality Reduction

3- Exploratory visualization of marker genes

4- **Clustering** / Hierarchies

5- **Differential Expression / Gene signature extraction**

6- Functional interpretation

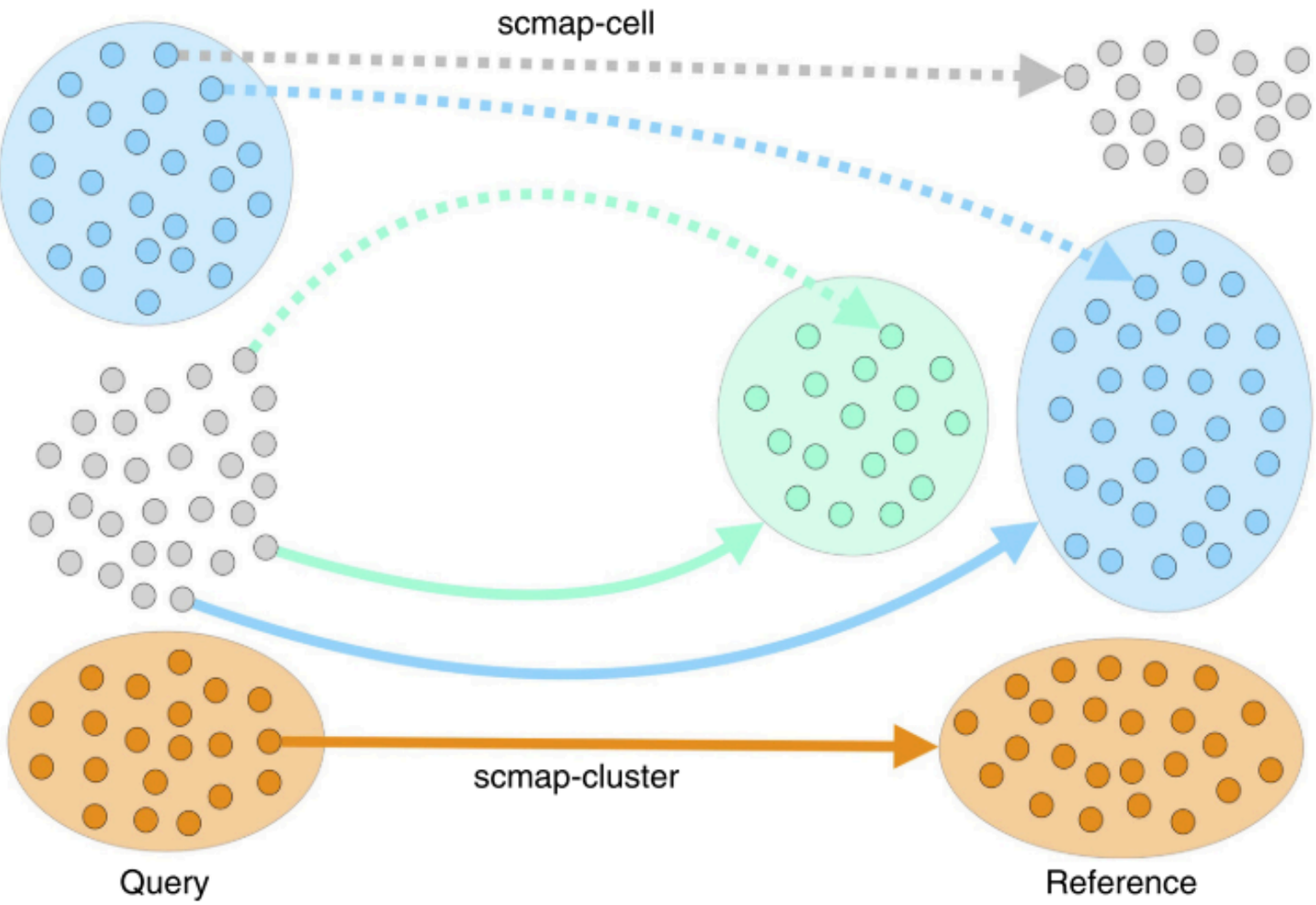
7- **A note on statistical robustness**

Today

8 - Batch Effect correction and data integration

9 - Single-cell matching across datasets

scmap cluster & scmap cell



5,000, or all genes. We calculated similarities by using the cosine similarity and Pearson and Spearman correlations, which are restricted to the interval [−1, 1] and are thus insensitive to differences in scale between data sets. We required that at least two of the similarities be in agreement, and that at least one be >0.7. If these criteria were not met, then *c* was labeled as "unassigned" to indicate that it did not correspond to any cell type present in the reference. For the approximate nearest neighbor search, which we refer to as scmap-cell, we carried out a form of *k*-nearest -neighbor classification with only cosine similarity. For a cell type to be assigned, we required that the three nearest neighbors have the same cell type and that the highest similarity among them be >0.5.

scmap: projection of single-cell RNA-seq data across data sets. Kiselev et al. Nature Methods 2018
http://www.nature.com/doifinder/10.1038/nmeth.4644

Thanks for your attention!
Twitter: @AntonioRausell
antonio.rausell@institutimagine.org