INTEGRATION

warning: statistics inside

INTEGRATION

Aggregation of data from different sources, and correcting for unwanted variation.

- mandatory reading:
  - Aaltonen et al., “Pan-Cancer Analysis of Whole Genomes.”
  - Ryu et al., “Integration of Single-Cell RNA-Seq Datasets.”
- suggested reading:
  - Korsunsky et al., “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony.”
  - Song, Chan, and Wei, “Flexible Experimental Designs for Valid Single-Cell RNA-Sequencing Experiments Allowing Batch Effects Correction.”

TWO TYPES OF INTEGRATION

HORIZONTAL INTEGRATION

combination of multiple data sources of the same type (modality) to increase sample sizes.

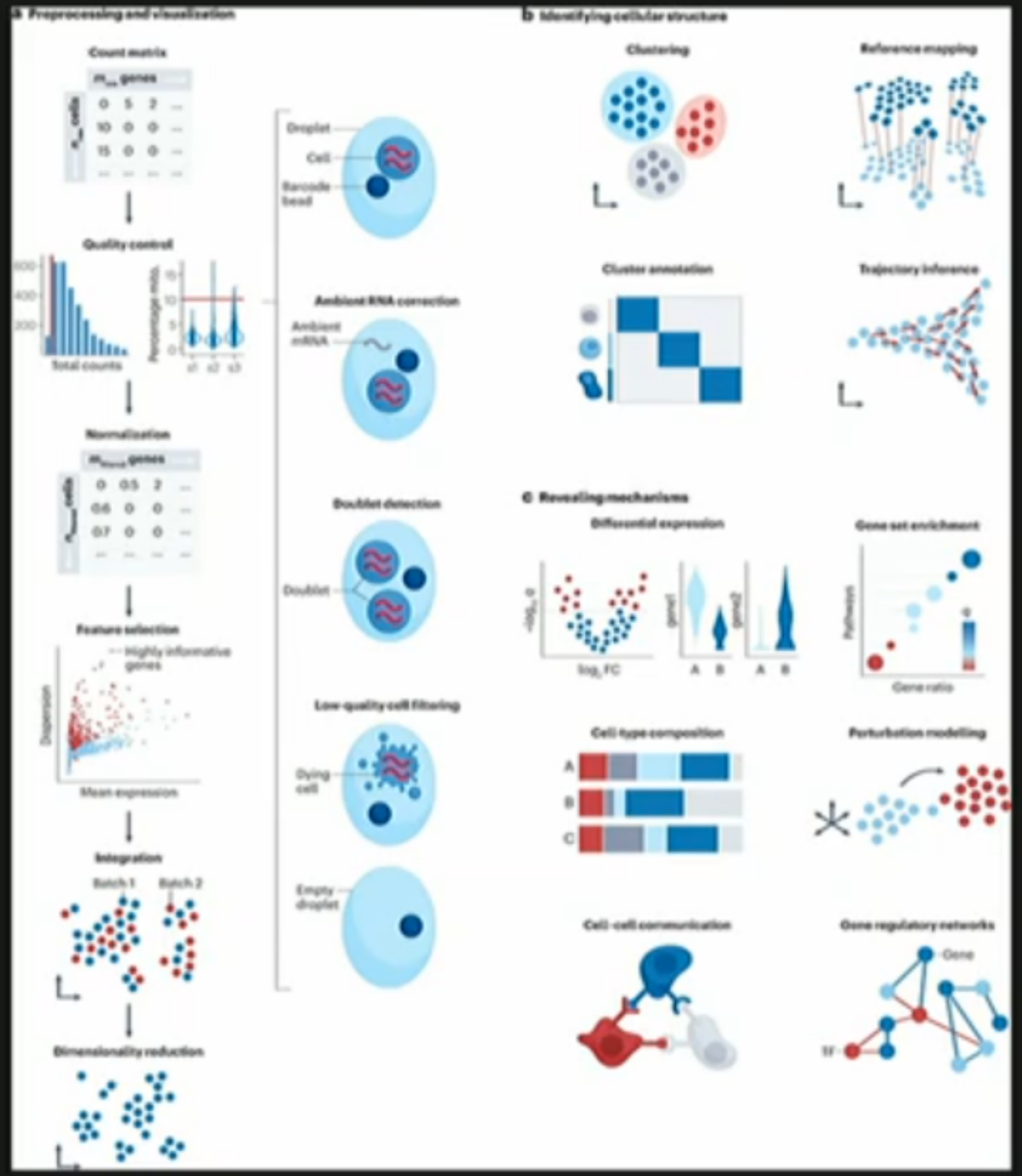
VERTICAL INTEGRATION / MULTIMODAL INTEGRATION

Combination of different types of data, to examine the same question from multiple points of view. e.g. chromatin accessibility + RNA expression + protein abundance

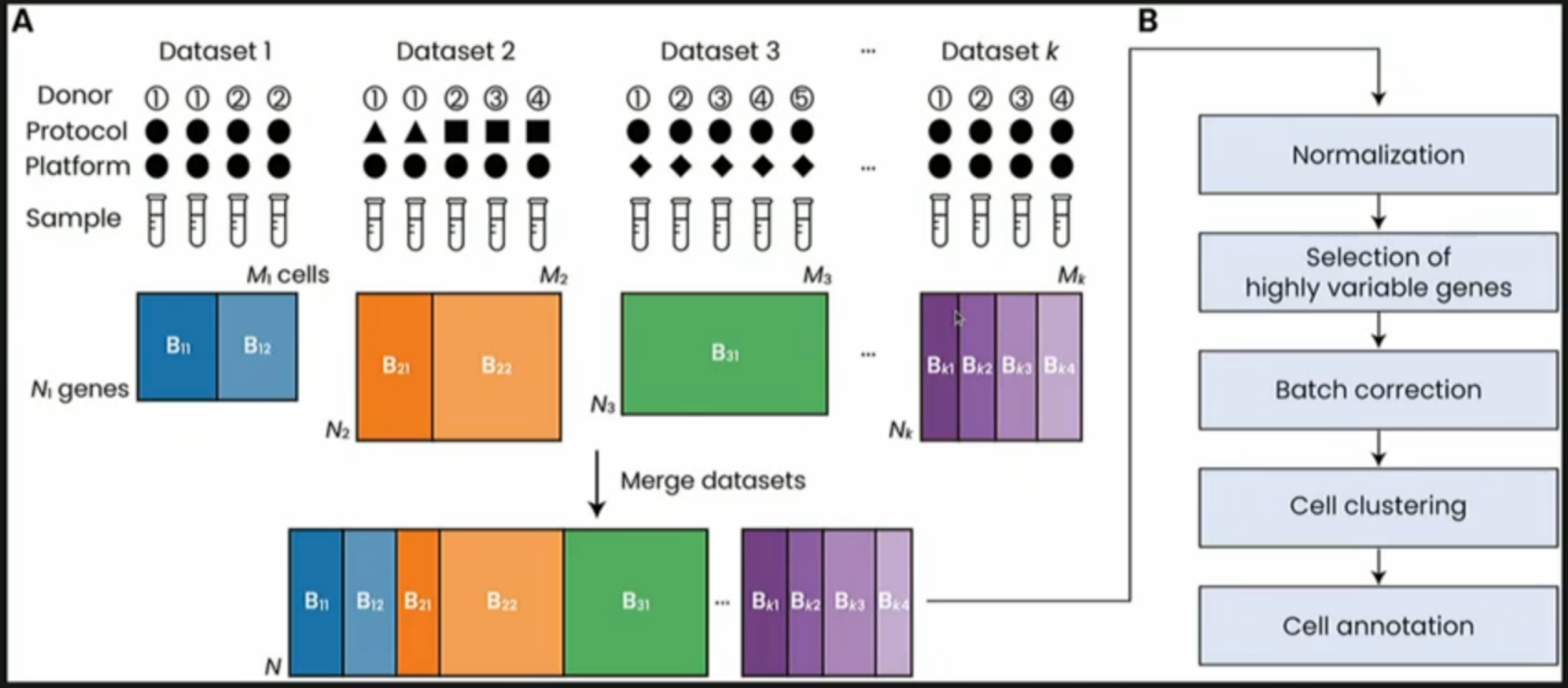
HORIZONTAL INTEGRATION <--

with a focus on scRNA-seq

A VERY BRIEF GUIDE TO SINGLE-CELL OMICS



BATCHES


A
Dataset 1
Dataset 2
Dataset 3
...
Dataset k
Donor
Protocol
Platform
Sample
M1 cells
M2
M3
Mk
N1 genes
N2
N3
Nk
B11
B12
B21
B22
B31
Bk1
Bk2
Bk3
Bk4
Merge datasets
N
B
Normalization
Selection of highly variable genes
Batch correction
Cell clustering
Cell annotation

BATCH EFFECTS

Any structural difference in expression between batches not caused by a variable of interest

ADVERSE EFFECTS

- Reduce statistical power by introducing noise, leading to false-negatives
- Introduce false-positives where batch effects are interpreted as meaningful
- Hurt the credibility of the experiment

FALSE POSTIVES

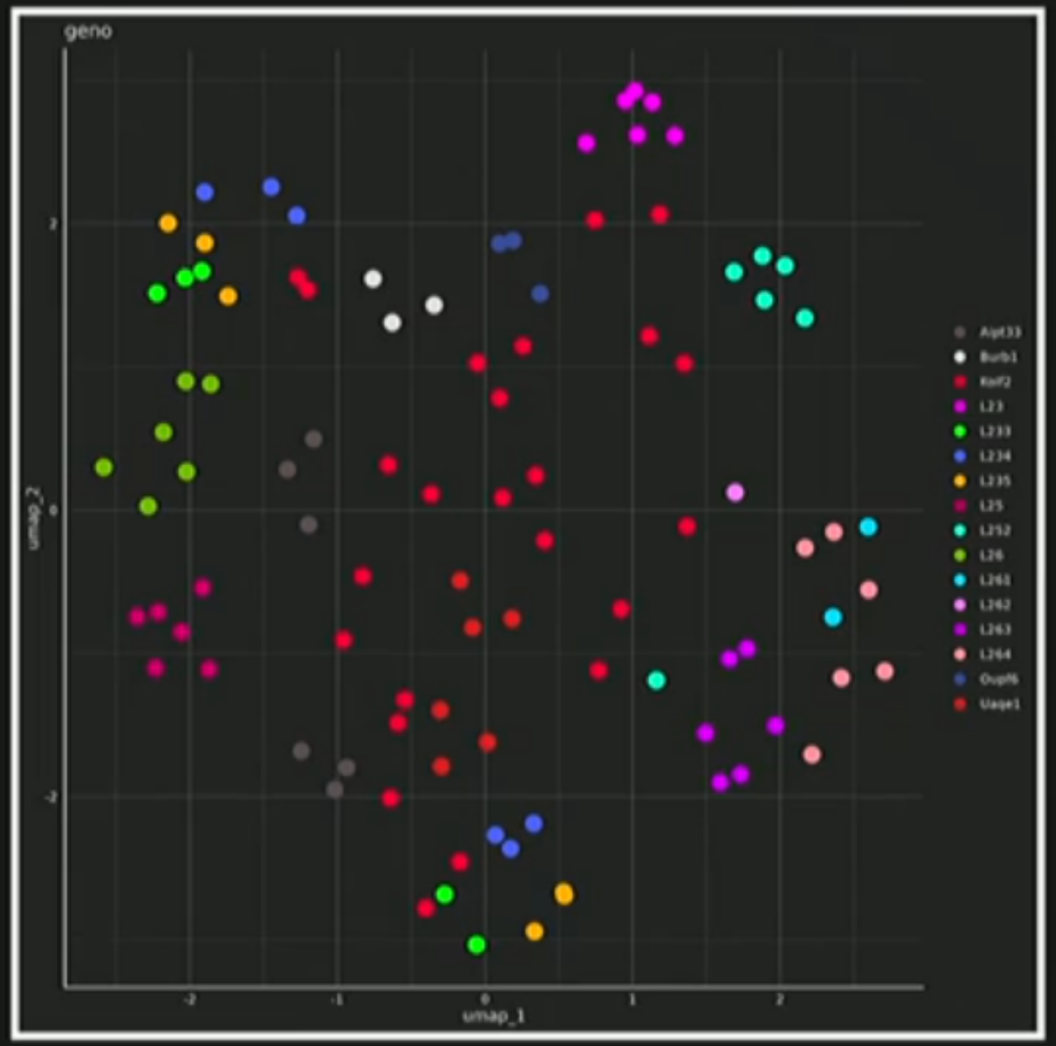
WHEN BATCH EFFECTS ARISE:

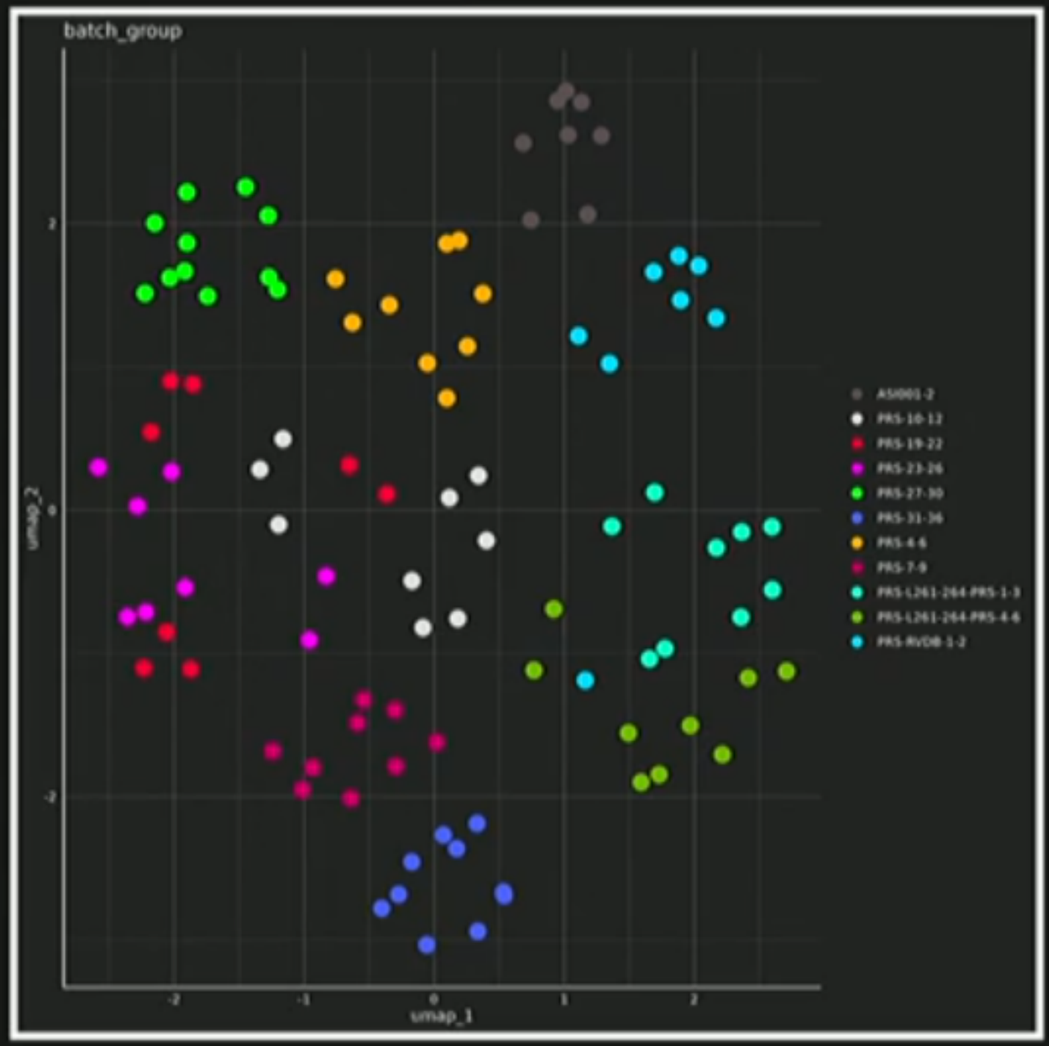
- When the experiment contains too many cells to fit a single sequencing machine.
- When data is sequenced on different days.
- When data is gathered from multiple participants.
- When combining multiple public datasets.
- When a lab technician sneezes.
- ...

ASSESSING BATCH EFFECTS IN OMICS DATA

- UMAP/PCA
- Differential cluster abundance
- Differential expression

ASSESSING BATCH EFFECTS - UMAP





EXPERIMENTAL DESIGN IS KEY

Limiting dependence between batches and variables of interest

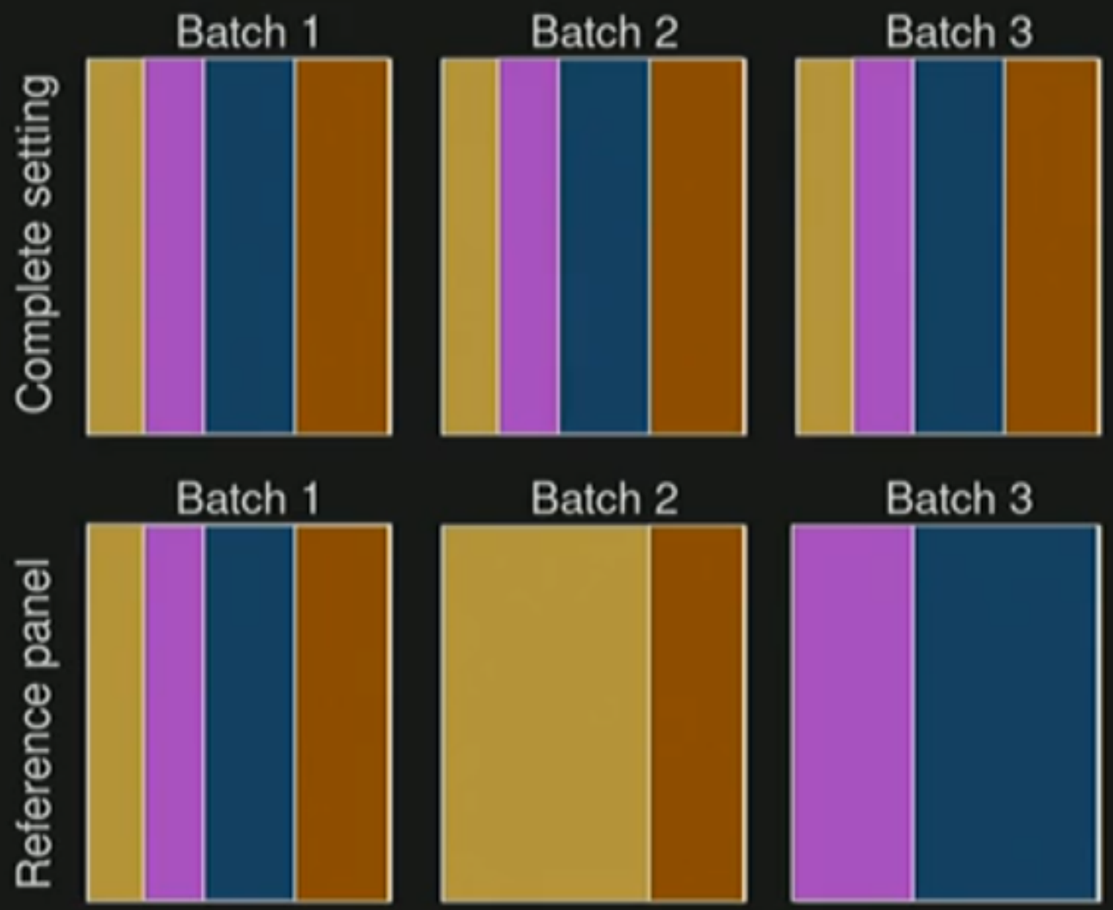
- Creates smaller batch effects, easier to correct for.
- Less loss of biological signal, when correction is needed.

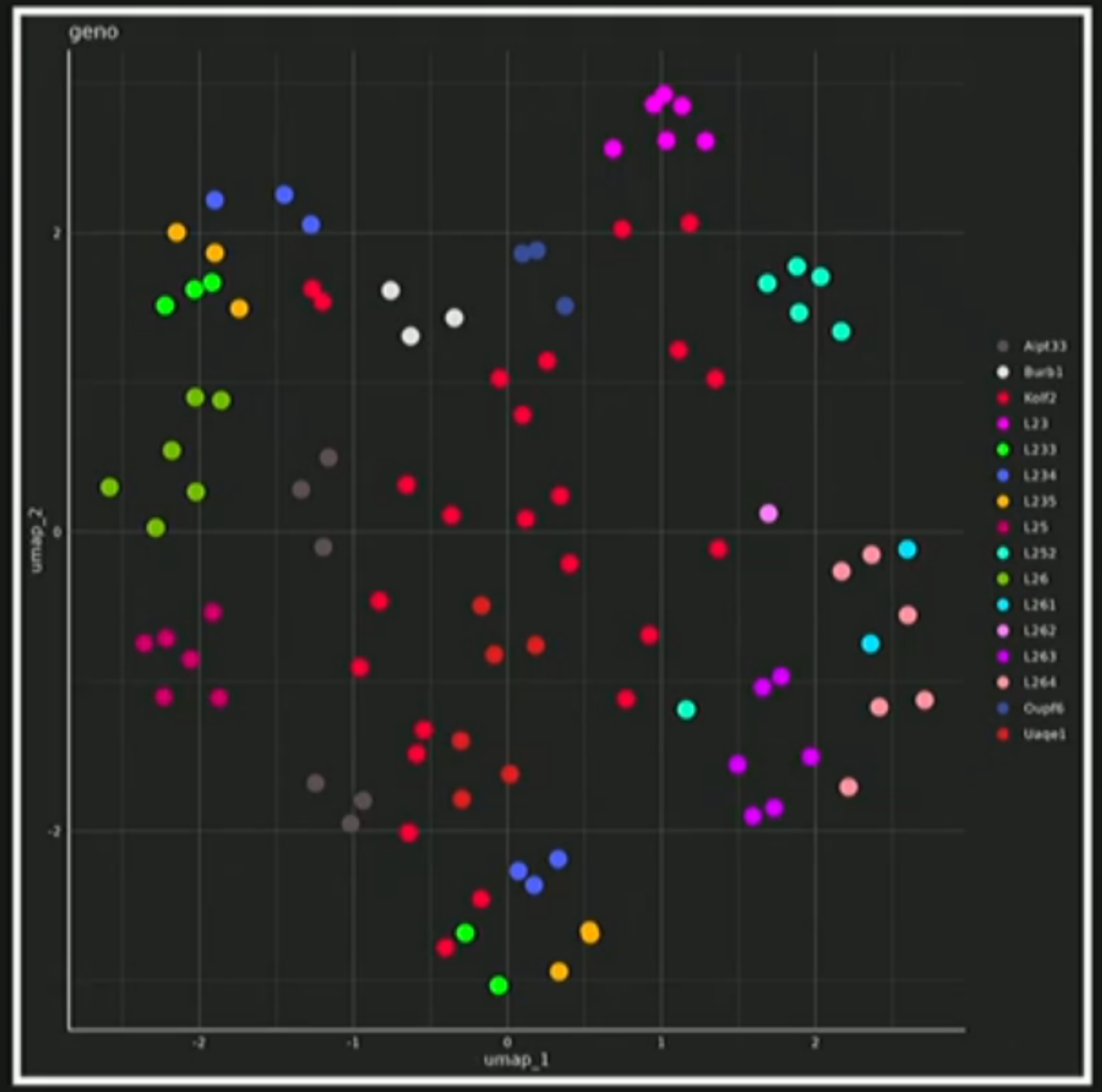
In simple terms: make sure every variable you are interested in is spread across variables you are not interested in.

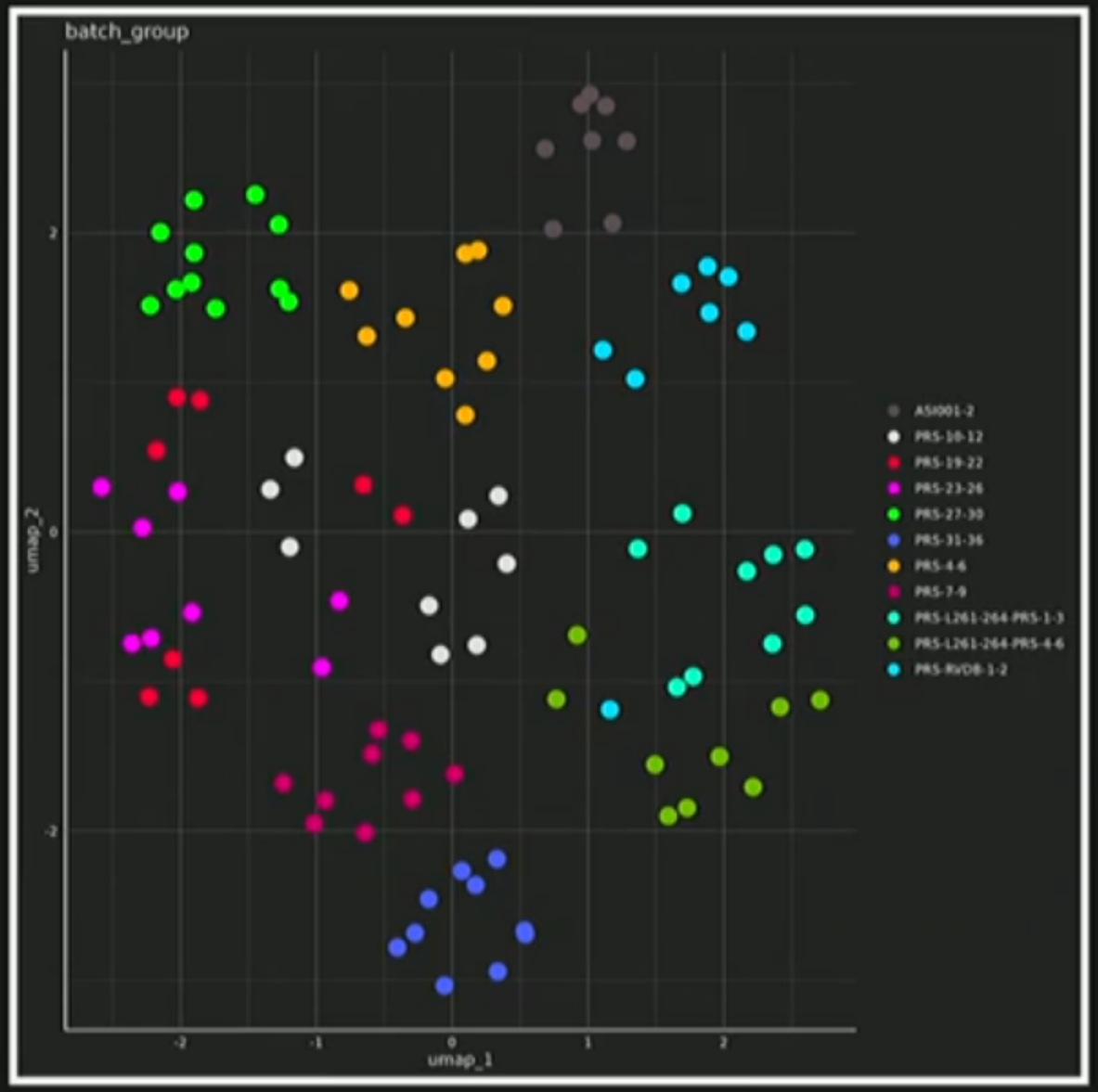
unfortunately often times bioinformatics are only consulted when this part of the process has passed, and data is generated.

THE WORST STUDY DESIGN

BETTER DESIGNS




geno
umap_2
umap_1
2
0
-2
-2
-1
0
1
2
Aipt33
Barb1
Kolf2
L23
L233
L234
L235
L25
L252
L26
L261
L262
L263
L264
Oupf6
Uaqe1

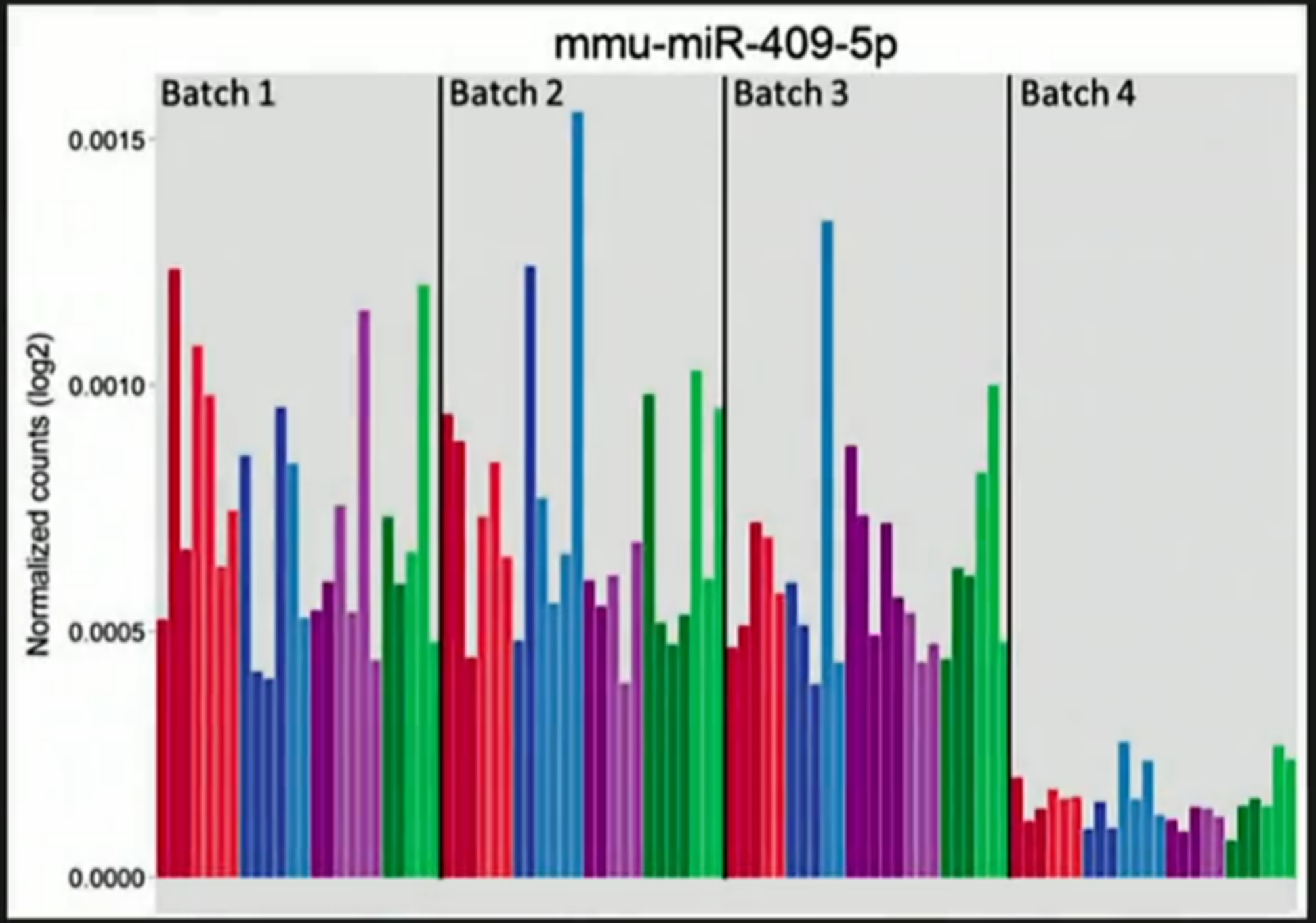

batch_group
umap_2
umap_1
2
0
-2
-2
-1
0
1
2
ASI001-2
PRS-10-12
PRS-19-22
PRS-23-26
PRS-27-30
PRS-31-36
PRS-4-6
PRS-7-9
PRS-L261-264-PRS-1-3
PRS-L261-264-PRS-4-6
PRS-RVDB-1-2

QUESTION

You want to do a meta-study of public AD data to examine the differences in microglia response to plaque load between patients, resilients and healthy controls. Choose 3 datasets to integrate into a new study with:

1. A complete setting design
2. A reference panel design
3. A reference condition design
4. A chain design
5. A confounded design

Study	AD	RES	CTRL
1	10000	10000	10000
2	10000	10000	10000
3	10000	10000	10000
4	0	10000	10000
5	10000	0	10000
6	10000	0	10000
7	10000	10000	0
8	0	0	10000
9	10000	0	0
10	0	10000	0


mmu-miR-409-5p
Batch 1
Batch 2
Batch 3
Batch 4
0.0015
0.0010
0.0005
0.0000
Normalized counts (log2)

A NOTE ON PRE-PROCESSING

When combining data it is best to start from raw. Any differences in the pre-processing are likely to cause significant batch effects.

INTEGRATION METHODS

MANY OPTIONS WITH KEY DIFFERENCES

- Correct count matrix or correct a dimensional reduction?
- Robust to multicollinearity?
- Aggressive or conservative?
- What metric or heuristic is optimized?

DIFFERENT METHODOLOGIES

- Regression models
- Linear decomposition models
- Similarity of cells based in reduced dimension space methods
- Similarity of clusters based in reduced dimension space methods
- Generative models with variational-autoencoders

INTEGRATION BY REGRESSION

Simple but effective when appropriate

- assumes linear relation between batch and counts
- assumes all cells are affected in the same way
- requires little to no multicollinearity
- alters the count matrix
- implemented in SCT/scanpy/seurat ...
- standard regression for normalized counts
- (zero inflated) negative binomial regression for raw counts

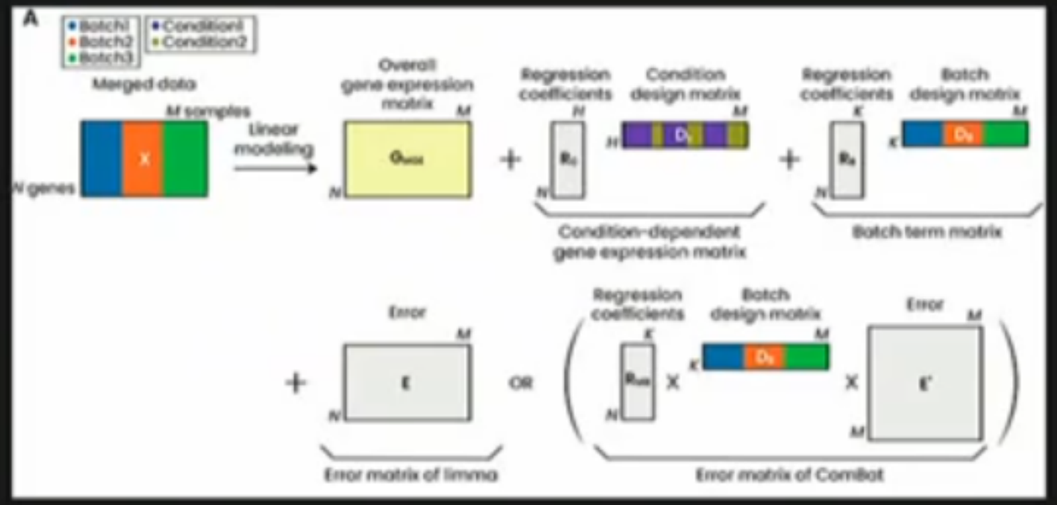
LINEAR DECOMPOSITION

aka matrix factorization. Mostly only used in bulk-RNA-seq integration

LIMMA

Solves a multivariate generalized linear regression including both variables of interest and batch effects

- assumes linear effects
- assumes all cells are affected in the same way
- requires little to no multicollinearity
- alters the count matrix



CELL-LEVEL SIMILARITY

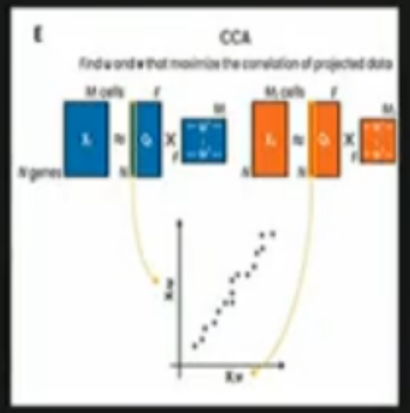
Methods which attempt to identify similar cells using various types of dimensional reduction or graphs.

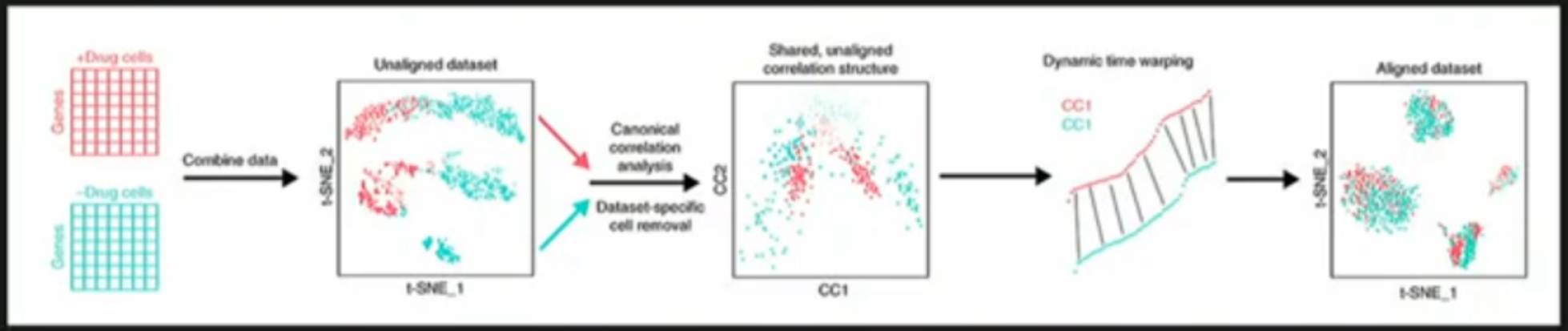
SEURAT-CCA

- handles sparsity of single-cell-seq by working in a dimensional reduction
- does not alter count data
- very popular

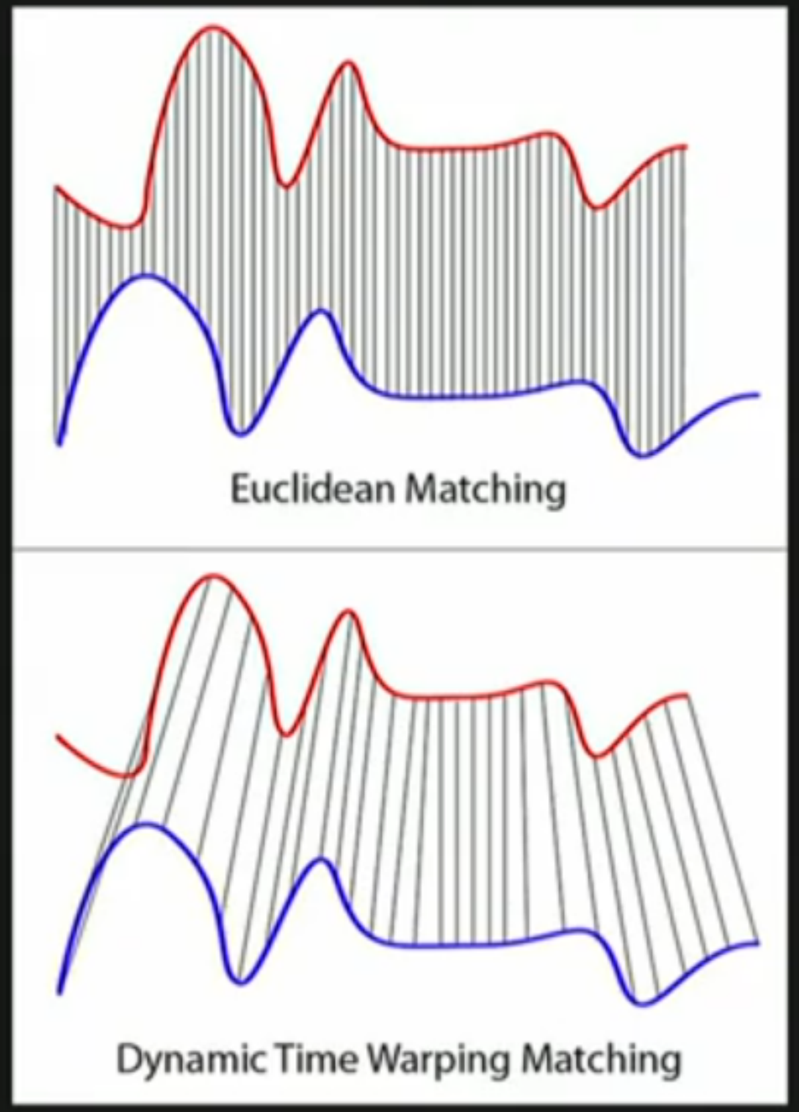
SEURAT-CCA

1. perform canonical-correlation analysis to generate a dimensional reduction with maximally correlated dimensions (but not alligned)
2. compute weighted average of genes which score the highest on these dimensions, i.e. meta-genes
3. linear rough alignment of 95% quantiles of metagenes
4. fine non-linear alignment using dynamic time warping
5. integrated analysis on aligned canonical basis vectors





DYNAMIC TIME WARPING

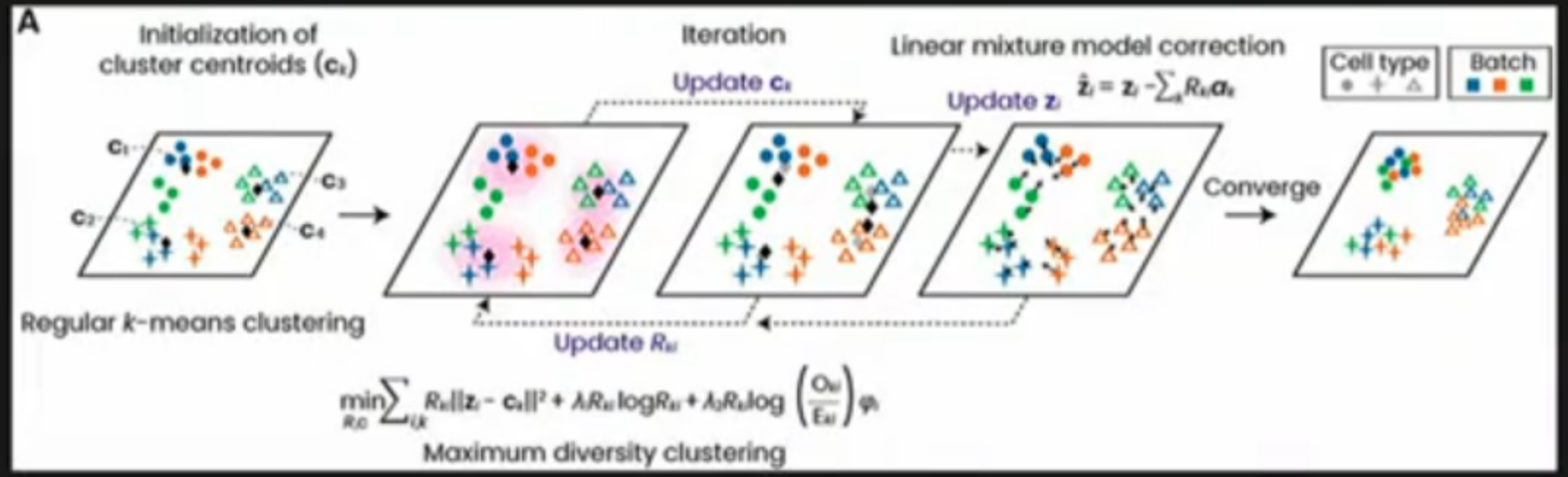


CLUSTER-LEVEL SIMILARITY

Methods which attempt to identify similar clusters of cells using various types of clustering.
Assumes cells within a cluster are biologically similar, suffer the same batch effects, and batches distribute evenly throughout the clusters

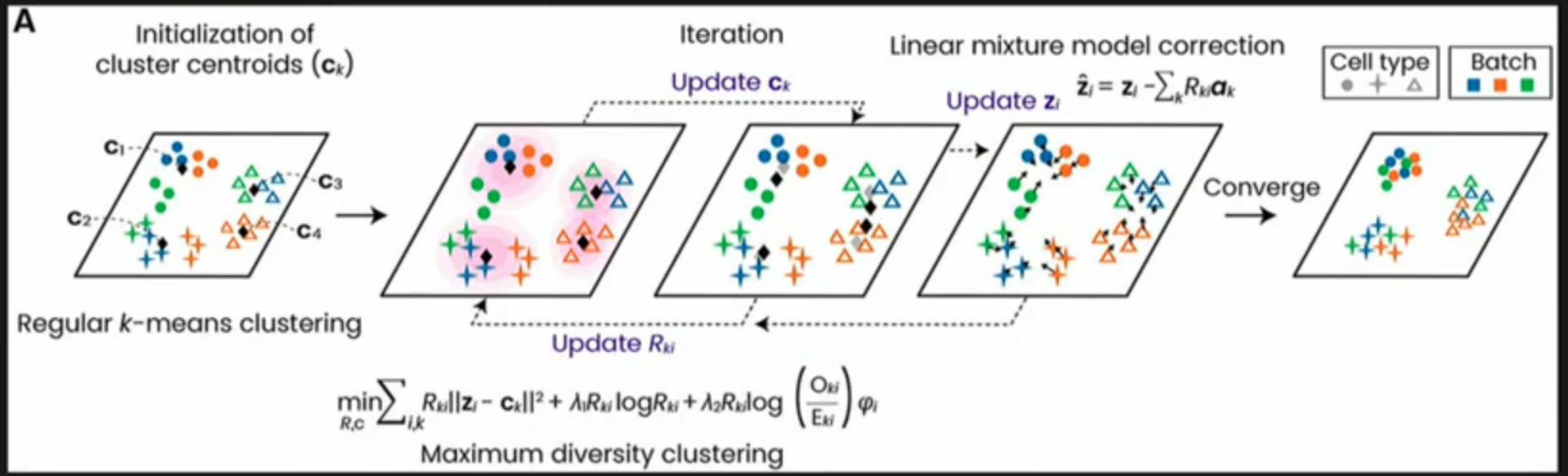
HARMONY

- the most popular option currently
- always among the best performing in review studies
- does not alter count data
- explicitly models different batch effects for each cluster -> more applicable to heterogeneous cell-types



HARMONY

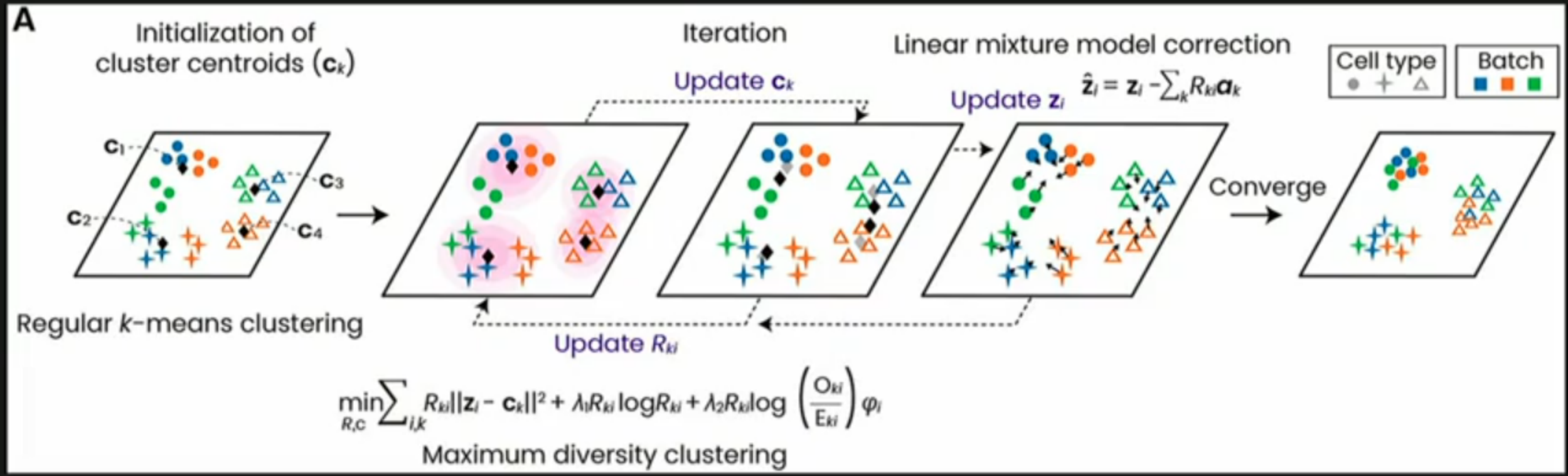
1. centroid generation & initial soft clustering
2. cycle of re-assigning cells to generate more diverse clusters
3. add correction to cells for each batch in each cluster
4. repeat until convergence



ADVANCED HARMONY

1. transform data to cosine distance
2. rounds of initial k-means clustering to get centroids
3. assign soft clusters to centroids by $R_{ki} \propto \exp(\frac{-||Z_i - Y_k||_2^2}{\sigma})$ where $\sum_k R_{ki} = 1$
4. calculate cluster diversity as the cross product of the one hot encoded design matrix and the cluster assignment matrix $R\phi^T$
5. assign diverse clusters by $R_{ki} \propto \exp(\frac{-(2(1-Y^T Z))}{\sigma})(\frac{E}{O})^\theta \phi$ a mini-batch of cells at a time, because moving cells will change diversity scores and centroids
6. recalculate centroids
7. 4 -> 6 until all cells are re-assigned
8. Approximately solve a mixture of experts model of the form $Z_d = \sum_k \beta_{0,k} + \beta_{1,k}\mathbb{1}_{(dataset=jurkat)} + \beta_{2,k}\mathbb{1}_{(dataset=half)} + \beta_{3,k}\mathbb{1}_{(dataset=293T)}$ where each cluster is an expert and each source dataset a predictive term by solving $W_k \leftarrow (\phi^* diag(R_k)\phi^{*T} + \lambda I)^{-1}\phi^* diag(R_k) Z_{orig}^T$
9. correct each cell by subtracting the batch specific term
10. 2 -> 9 K times

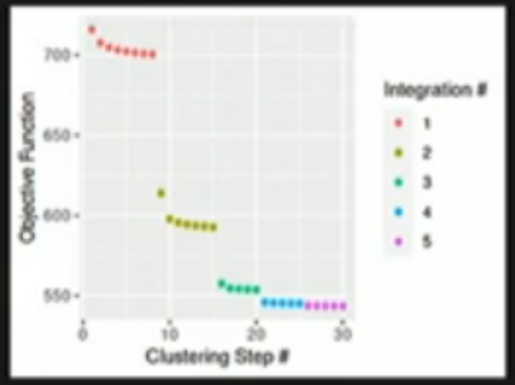
HARMONY - PARAMETERS



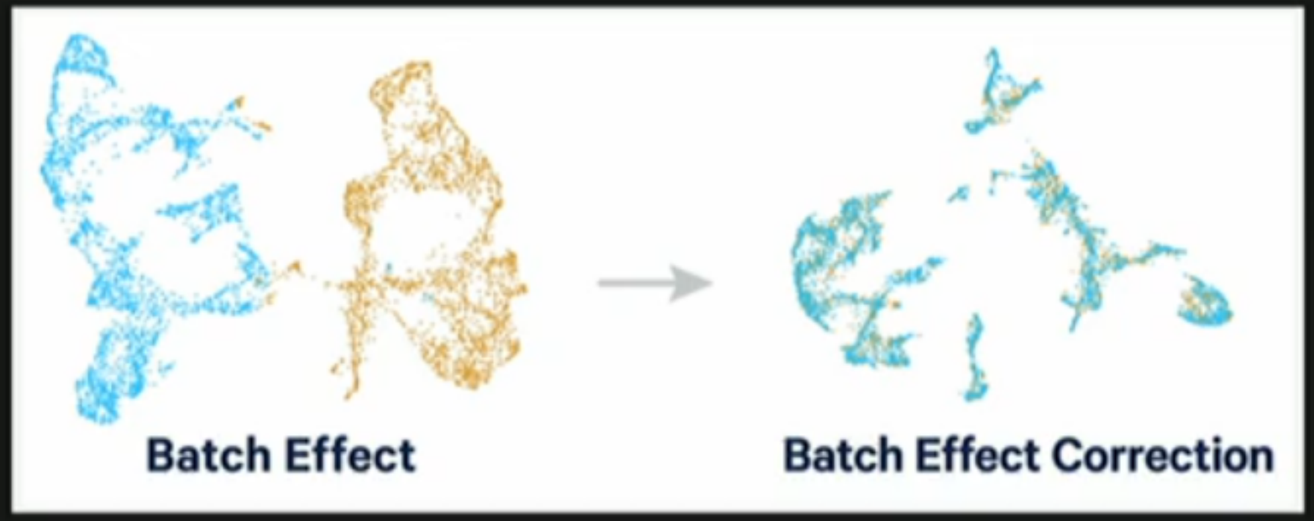
- softness of clustering
- penalty for dependence
- number of clusters
- minimum cells per cluster
- maximum number of iterations
- convergence limit

REAL LIFE USE OF HARMONY

```
RunHarmony(So,"batch", plot_convergence = TRUE, nclust = 50, max_iter = 10, early_stop = T)
```

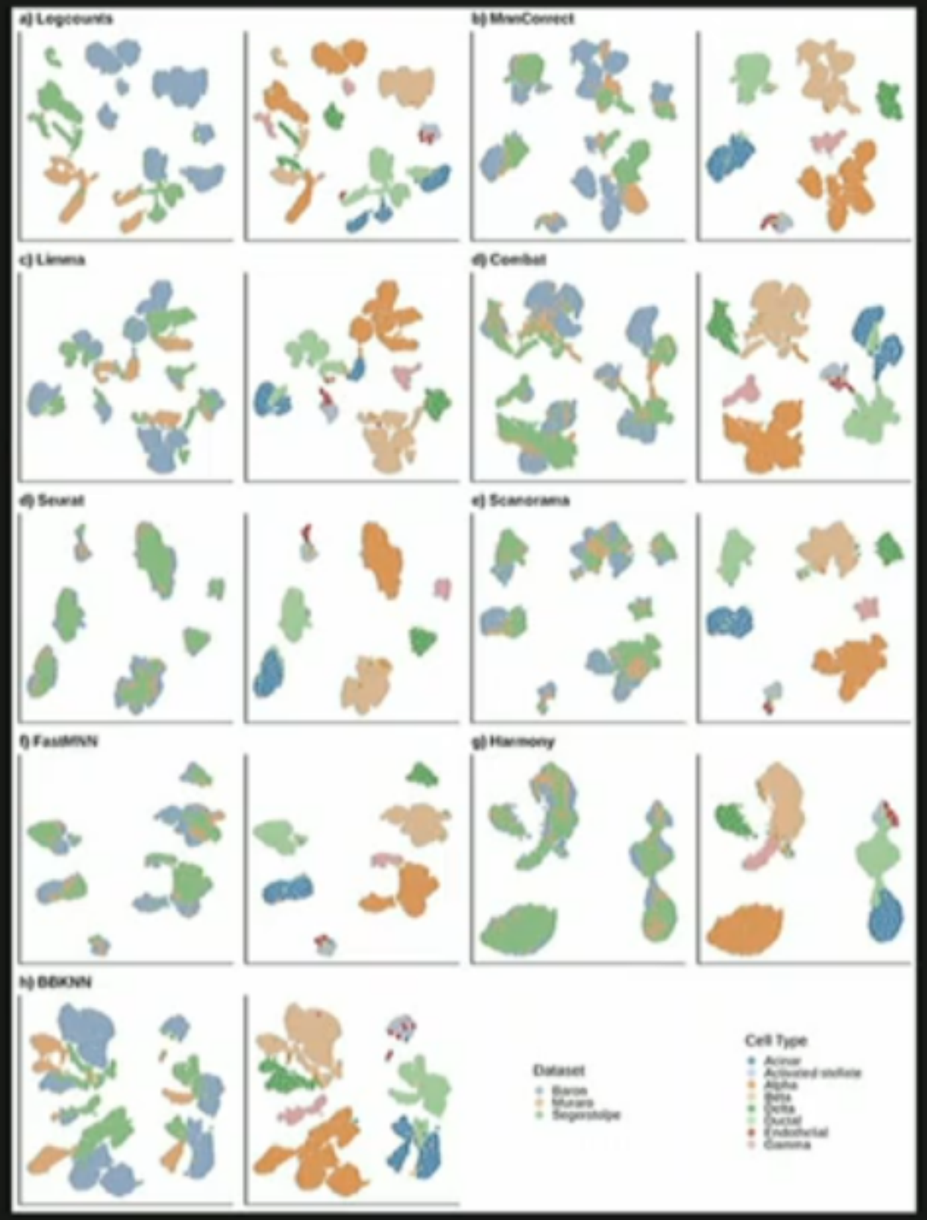


SUCCESS



80% of the time

ASSESSING INTEGRATION PERFORMANCE



WHEN INTEGRATION FAILS

A high level understanding of integration methods becomes key for the modern bio-informatician

- understand which parameters need tuning given the data
- try more appropriate algorithms for your data
- ensure pre-processing was comparable

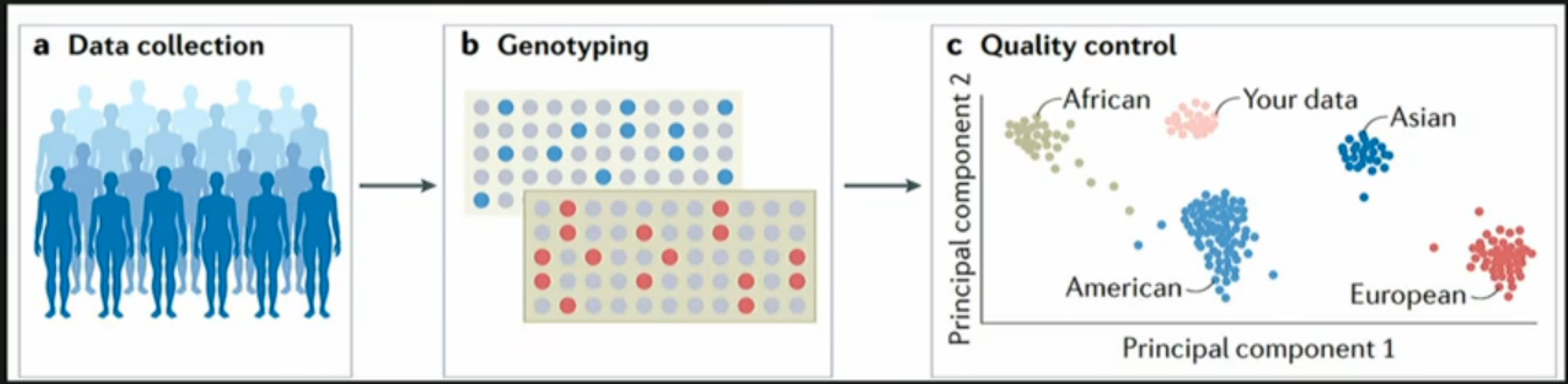
CORRECTION VS CONSERVATION

- More aggressive correction can lead to loss of biological signal and false negatives.
- Less aggressive correction may cause noise and false positives.
- The severity of this trade-off is proportional to the dependence between what you are interested in and what you are trying to correct for: see study design.

COMMON PITFALLS

- over-correcting data
- under-correcting data
- performing DE on uncorrected data after integration

INTEGEGRATION IN SNP CALLING



Assume you try to associated variants with Alzheimer's Disease, and you sample patients from a cohort of both European & Asian individuals:

Population	Alzheimer's	Healthy
European	100	10
Asian	10	100

what will happen?

ASKING THE WRONG QUESTION CAN BE DANGEROUS:

cases and controls should be matched by ancestry to avoid confounding; for example, a GWAS for chopstick use where cases are defined as ‘using chopsticks regularly’ and controls as ‘not using chopsticks’ would likely result in cases being drawn more often from an East Asian population than controls.
Read: Uffelmann et al.

HOW TO DEAL WITH THIS?

- Careful selection of your cohort (maybe focus on one population)
- Careful recording of relevant parameters
- Exclude outliers
- Identification of relevant stratification (PCA)
- Use (confounding) principal components in as covariates

PANCANCER PAPER

Article

Pan-cancer analysis of whole genomes

https://doi.org/10.1038/s41586-020-1969-6

Received: 29 July 2018

Accepted: 11 December 2019

Published online: 5 February 2020

Open access

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium

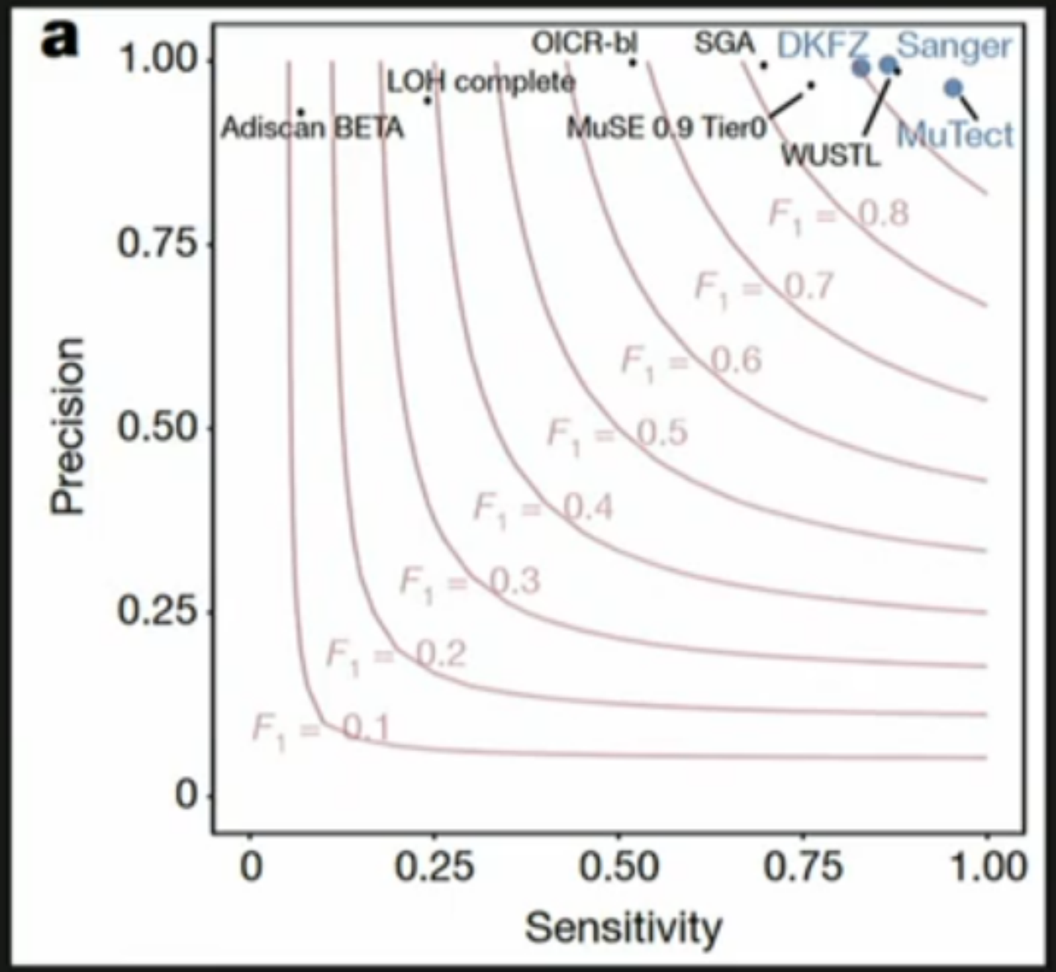
Cancer is driven by genetic change, and the advent of massively parallel sequencing has enabled systematic documentation of this variation at the whole-genome scale[1–3]. Here we report the integrative analysis of 2,658 whole-cancer genomes and their matching normal tissues across 38 tumour types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). We describe the generation of the PCAWG resource,

PANCANCER PAPER

- Integrated analysis of 2,658 cancer genomes across 38 tumor types
  - Whole genome shotgun
    - SNV
    - Indels
    - Copy Number Variants
    - Structural Variation
    - Somatic retro-transposition events
    - Mitochondrial DNA mutations
    - Telomere lengths
    - Analysis of germline variants
- RNAseq -> transcriptional changes

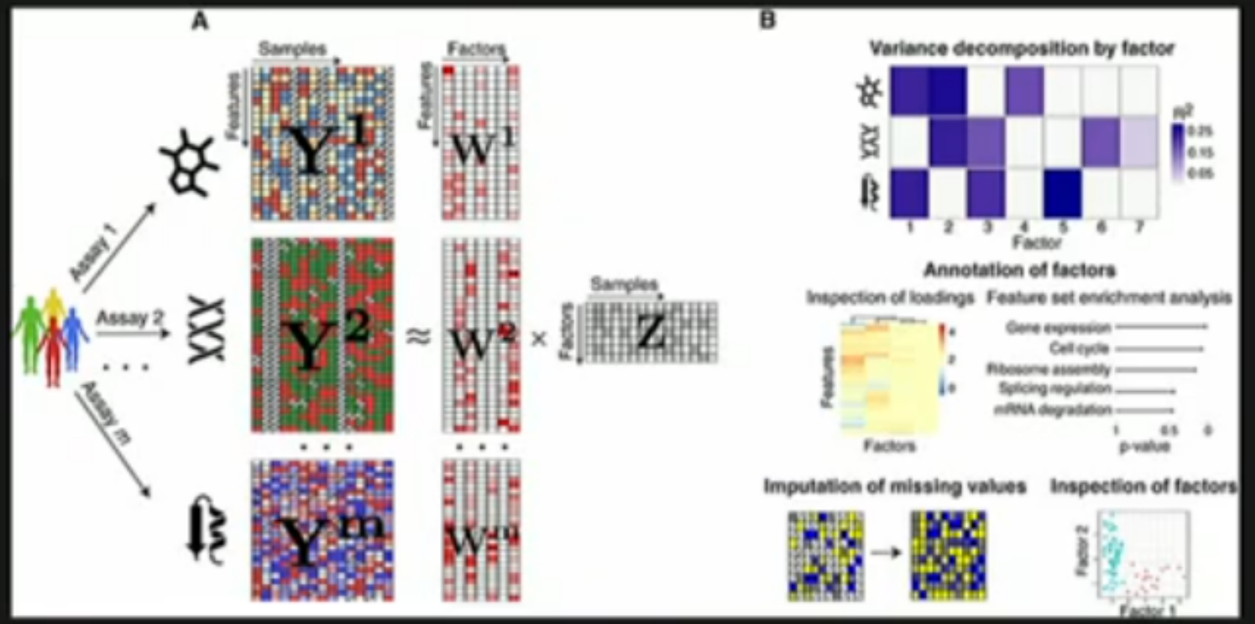
VALIDATION OF SNP CALLING ALGORITHMS

- compared 13 algorithms
- Confirmed in 50 individuals
- Merge results of top 3 into one final call



VERTICAL INTEGRATION WITH MOFA

- Linear decomposition method
- takes a multi modal input
- scores each feature from each modality on a factor
- scores every sample on these multi-modal factors
- annotate factors based on feature loadings
- correlate these multi-modal factors with outcomes



KEY TAKE-AWAYS

- Controlling for batch effects starts with the experimental set-up
- Integration is an unsolved issue
- Integrating data is as much art as science and requires understanding of the algorithms used
- Integration methods can be very complicated
- A high level understanding of methods will often be enough to determine if they fit your data
- There are many methods with equally many different assumptions
- There exists a trade-off between integration and loss of biological signal
    - there are no free lunches