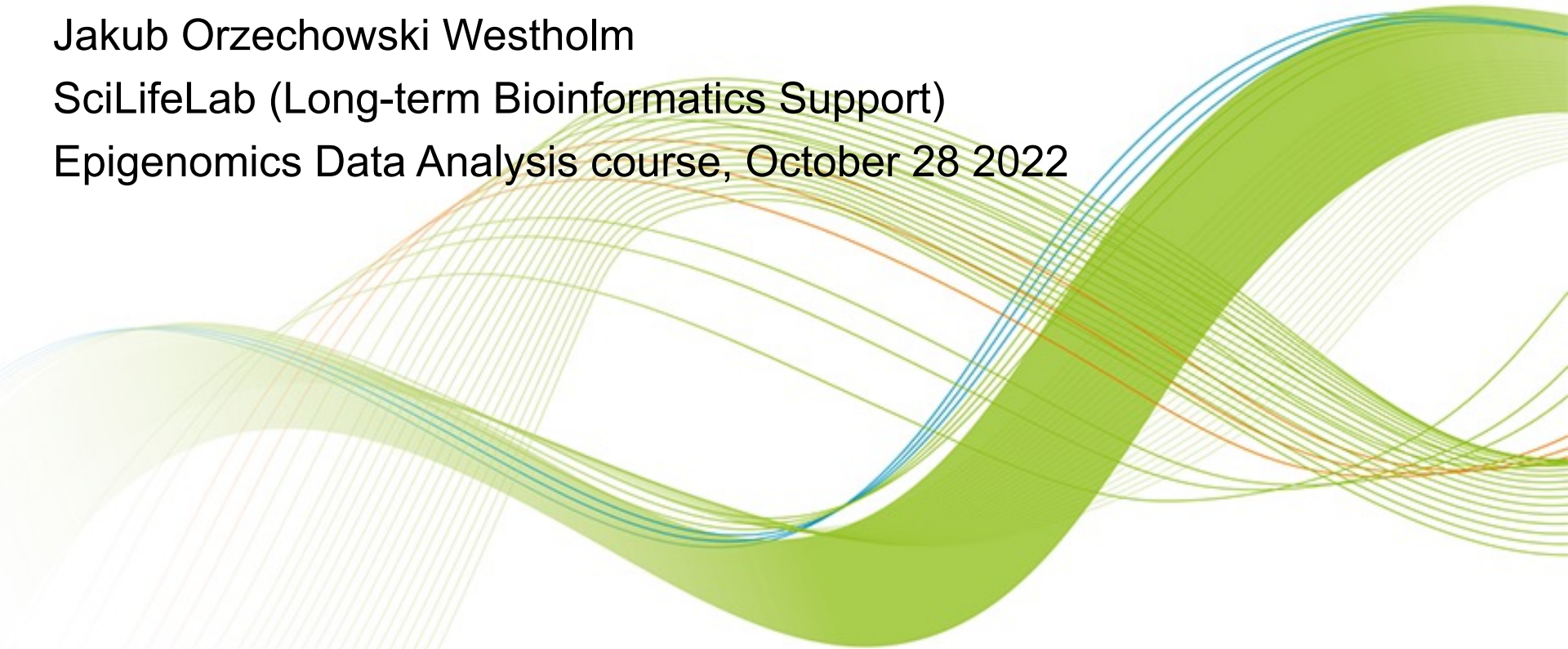
SciLifeLab

Integration of different types of genomics data

Jakub Orzechowski Westholm
SciLifeLab (Long-term Bioinformatics Support)
Epigenomics Data Analysis course, October 28 2022



What is data integration

SciLifeLab

- Combine data from different studies and different technologies, e.g.
  - RNA-seq
  - ATAC-seq
  - DNA-methylation
  - Spatial transcriptomics
  - Proteomics
  - Metabolomics
  - ..
- The goal: Learn something you couldn't learn from looking at each individual data set independently

Some data integration applications

SciLifeLab

- See how different data sets correlate, e.g. TF binding vs gene expression, two TF binding data sets, gene expression vs open chromatin (descriptive analysis).
- Learn something about one data set, from looking at another (common for single cell data, e.g. cell type composition, label transfer, deconvolution).
- Find important patterns in the data (unsupervised learning).
- Use several data sets to predict phenotype (supervised learning).
- …

This talk will be a pretty general discussion on what data integration can be used for, and some methods to do this.

How do we want to combine data?

SciLifeLab

- Combine data sets with the same features in different samples.

	Sample 1	Sample 2	…	Sample M
Gene 1				
Gene 2				
..				
Gene N				

	Sample M+1	Sample M+2	…	Sample M+K
Gene 1				
Gene 2				
..				
Gene N				

Remove batch effects

- Combine data sets with different features in the same samples.

	Sample 1	Sample 2	…	Sample M
Gene 1				
Gene 2				
..				
Gene N				

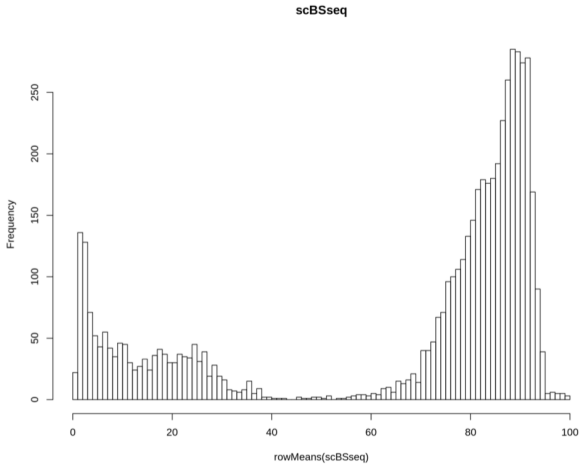
	Sample 1	Sample 2	…	Sample M
CpG 1				
CpG 2				
..				
CpG L				

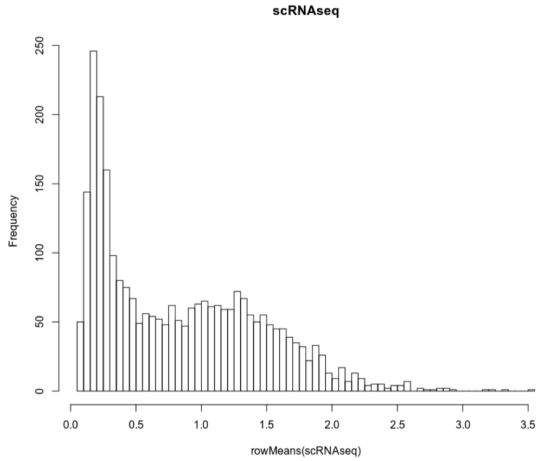
Makes some analysis harder

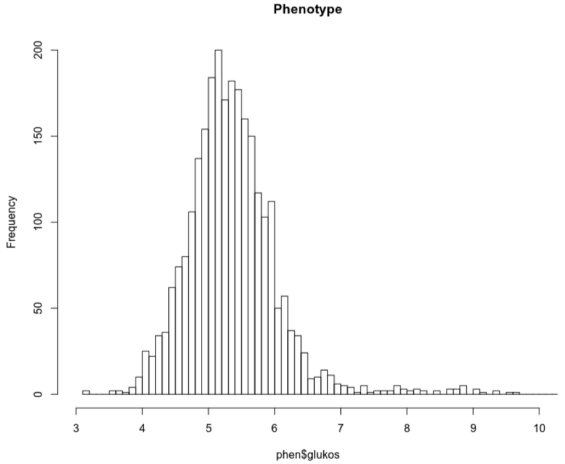
Combine features: Things to consider

SciLifeLab

- Not trivial how to combine data from different assays
  - Different number of features
  - Different biases
  - Different distributions





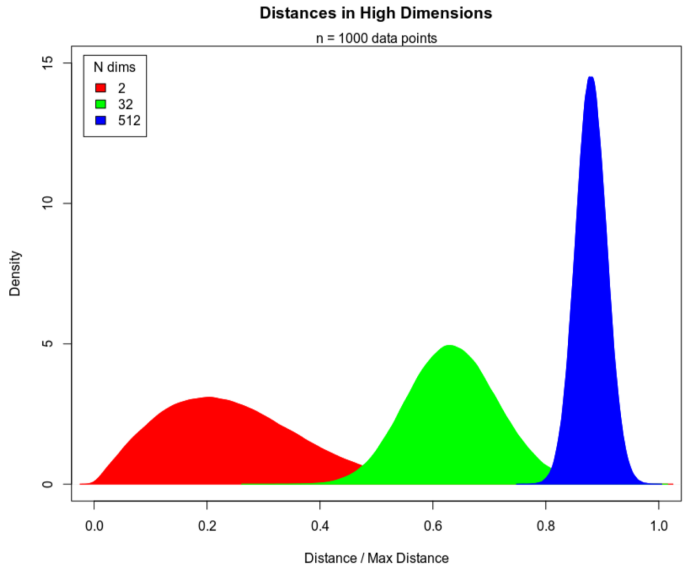


(Nikolay Oskolkov)

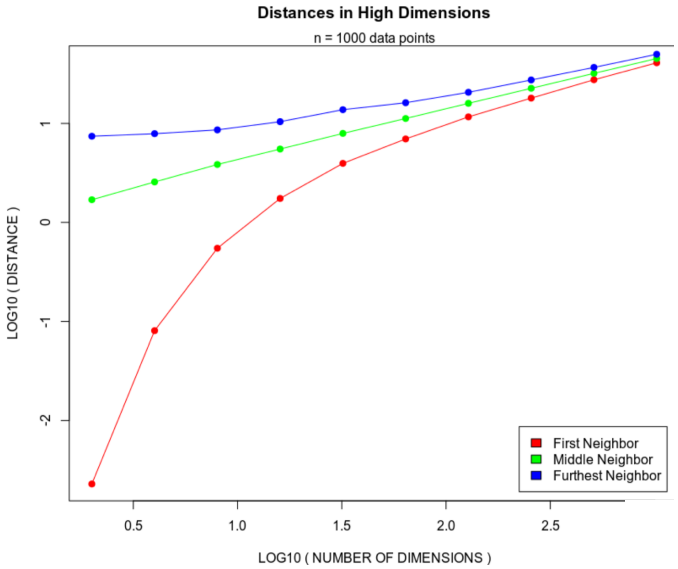
“Curse of dimensionality”

SciLifeLab

- Having too many features compared to samples (P>N) will cause problems for modeling and prediction.
- Too many features/dimensions makes clustering hard.



Data points end up far from each other and equidistant from each other in high dimensions.



The differences between closest and furthest data point neighbors disappears in high-dimensional spaces – can’t cluster.

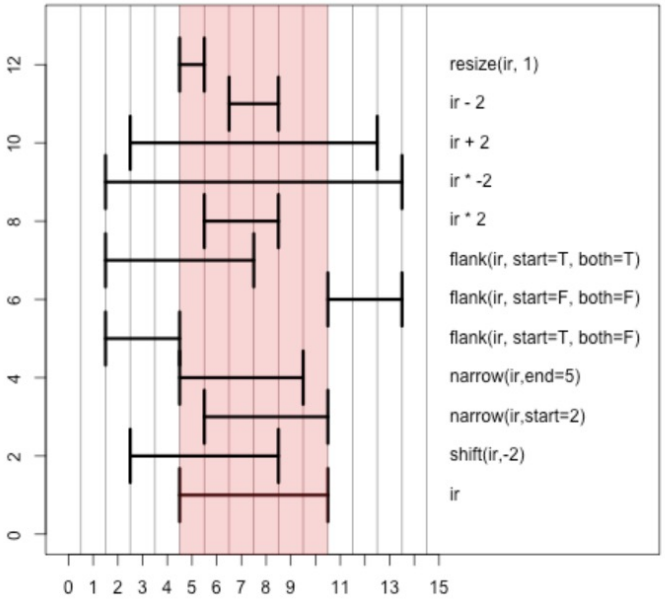
(Nikolay Oskolkov)

Descriptive analysis/Comparing data sets

Integration of genomic ranges

SciLifeLab

- Many types of data can be represented as regions across the genome
  - Genes
  - Promoters
  - CpG islands
  - Open chromatin
  - Transcription factor binding sites
  - SNPs

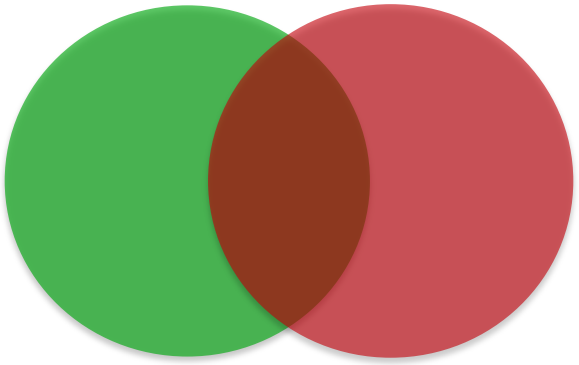


- R package to manipulate genomic ranges, **`GRanges`**.
  - Underlies a lot of tools in R.
  - We will look at **`GRanges`** in the exercise.
  - Bedtools has similar functionality

Integration of genomic ranges II

SciLifeLab

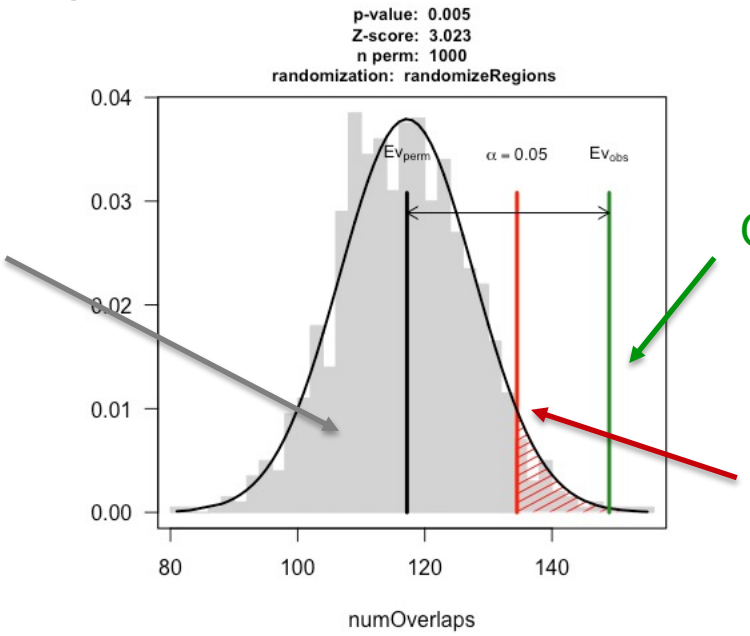
- A common application is to test if the overlap between two data sets is statistically significant (e.g. greater than expected by chance).
  - If the the binding sites of two transcription factors overlap
  - If a protein preferably binds around transposable elements, promoter regions, CpG islands etc.
  - …
- We will look at this in an exercise today.



Integration of genomic ranges III

- Many ways to do this, many ways to model what is expected by chance.
  - Theoretical (fast) vs empirical (many randomizations, slow)
  - Takes region size into account?
  - Takes distances between regions into account?
  - Takes linkage disequilibrium into account?
  - Do the lengths of the overlaps influence the result?
- This can have a big impact on results.

Overlaps from randomized data

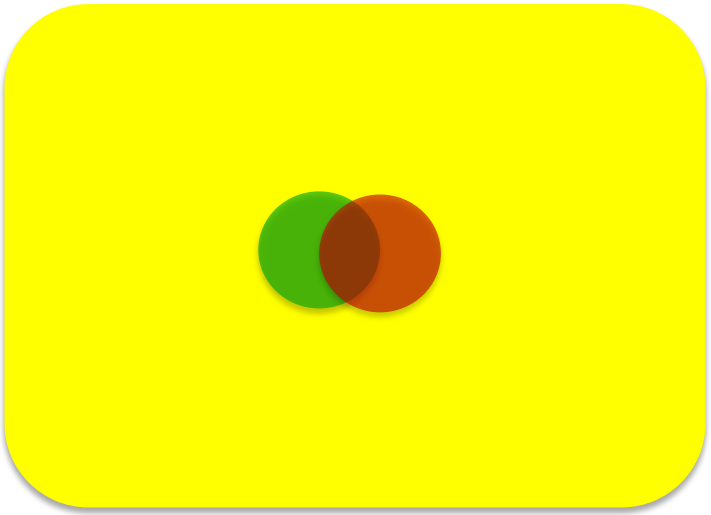


Observed overlaps

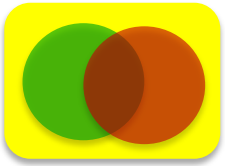
P=0.05, from randomized data

Integration of genomic ranges IV

- The most important difference between methods is the null distribution, i.e. how much overlap do we expect by chance?
- This depends a lot on the choice of universe, e.g. the whole genome or just promoters.



Big universe (e.g. whole genome)
→ significant overlap

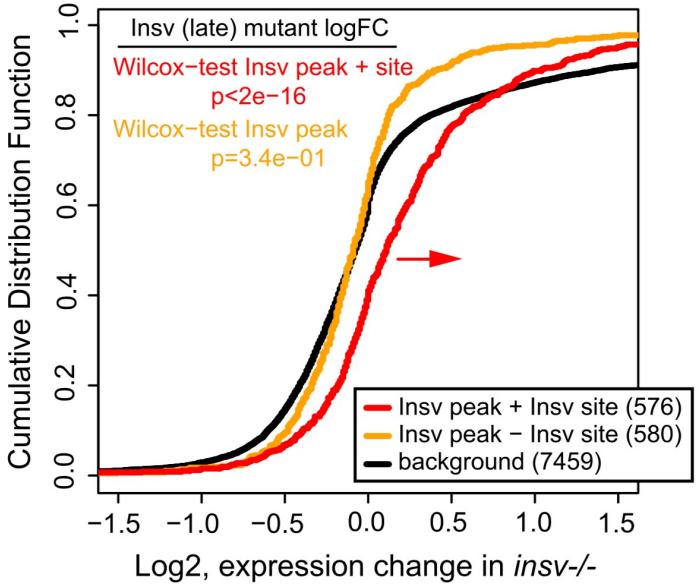


Small universe (e.g. promoters)
→ same overlap is not significant

Combining TF binding and gene expression

SciLifeLab

- A common application is to combine genomic regions (e.g. ChIP-seq peaks) and gene expression data, to see if the expression of genes near the region changes. For this we need
  - Annotation of each region, the nearest gene. (This might not be optimal..)
  - Gene expression data, processed in some useful format (TPM, log ratio etc.)



(Dai et al. 2013, Genes & Development)

Integration of single cell data

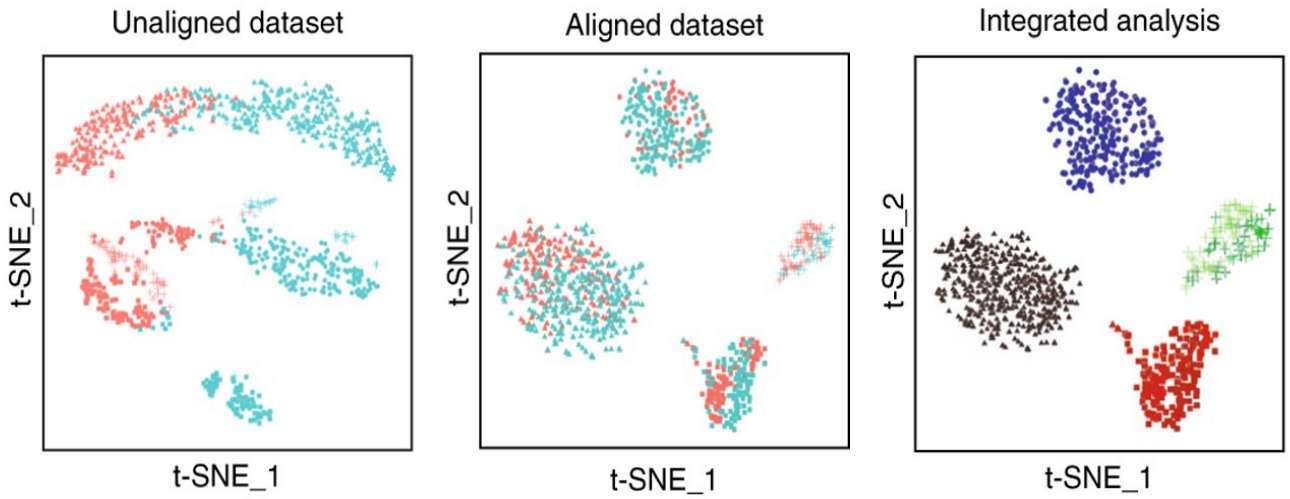
Integration of single cell data

SciLifeLab

- It is often useful to integrate different single cell data sets. Could be
  - from different batches
  - from different studies
  - your data vs public atlas
  - RNA-seq vs ATAC-seq
- Often: Look at the same features, but combine cells.
- Some version of this is used in many/most single cell studies.

Integration of single cell data, II

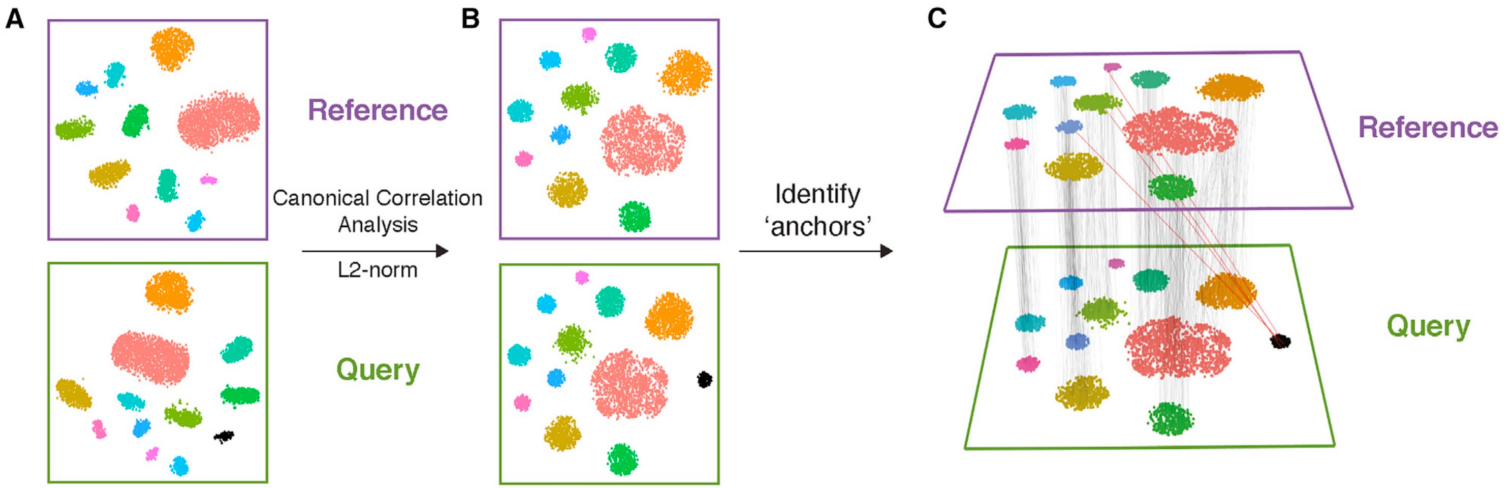
- Goal: Integrate data so that cells from different experiments (and –omics) can be analyzed together.
  - Make sure measurements are on the same format
  - Align data sets
  - Label transfer: Use cell type annotations from one data set (e.g. RNA-seq), to annotate another data set (e.g. ATAC-seq).



(Butler et al. 2018, Nature Biotechnology)

Data integration with Seurat

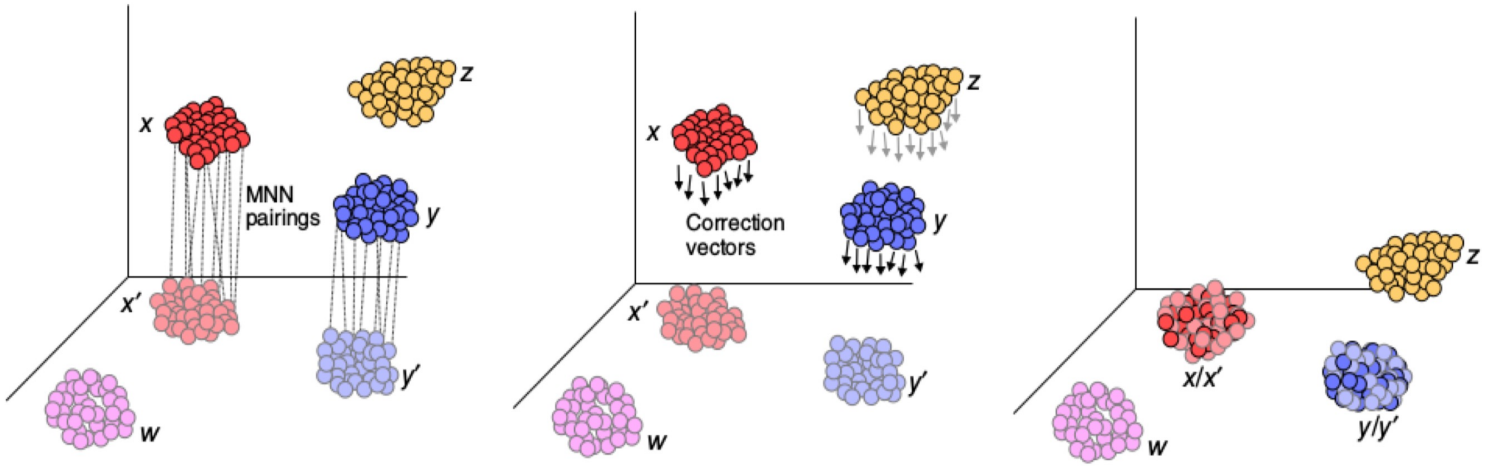
- There are many methods for integrating single cell data, but one of the more used is **`Seurat`**.
- How Seurat integrates two data sets, A and B, without getting too technical:
  1. Normalize both data sets
  2. Use CCA method to project both data sets to a common space
  3. Find “anchors”, e.g. pairs of cells from A and B that are mutual nearest neigbors (MNNs)
  4. Filter the MNNs, so we only keep high confidence pairs.



(Stuart et al. 2019, Cell.)

Data integration with Seurat, II

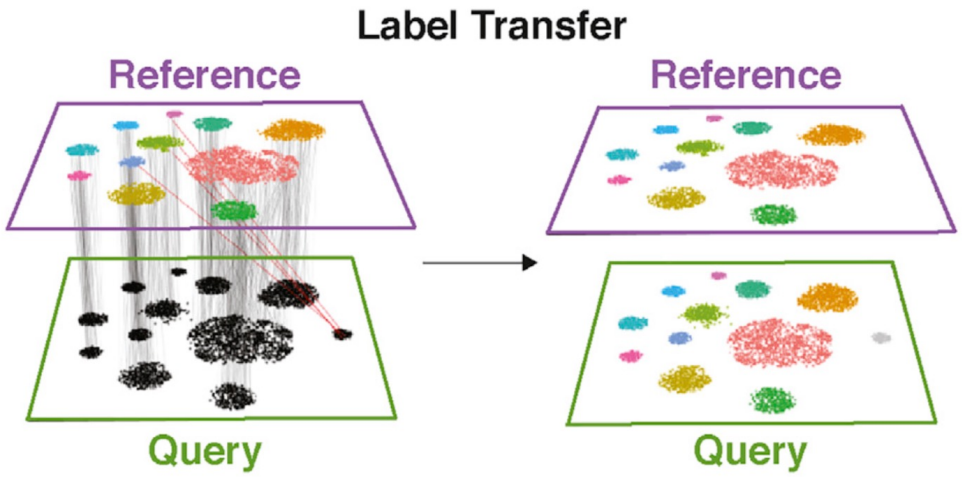
5. Get a correction vector for each anchor, which is the difference between the two cells in the pair.
6. Use the correction vectors to move the cells from data set A



(Haghverdi et al. 2018, Nature Biotechnology.)

Data integration with Seurat, III

7. Use cell type labels from B to assign cell type to cells in A.

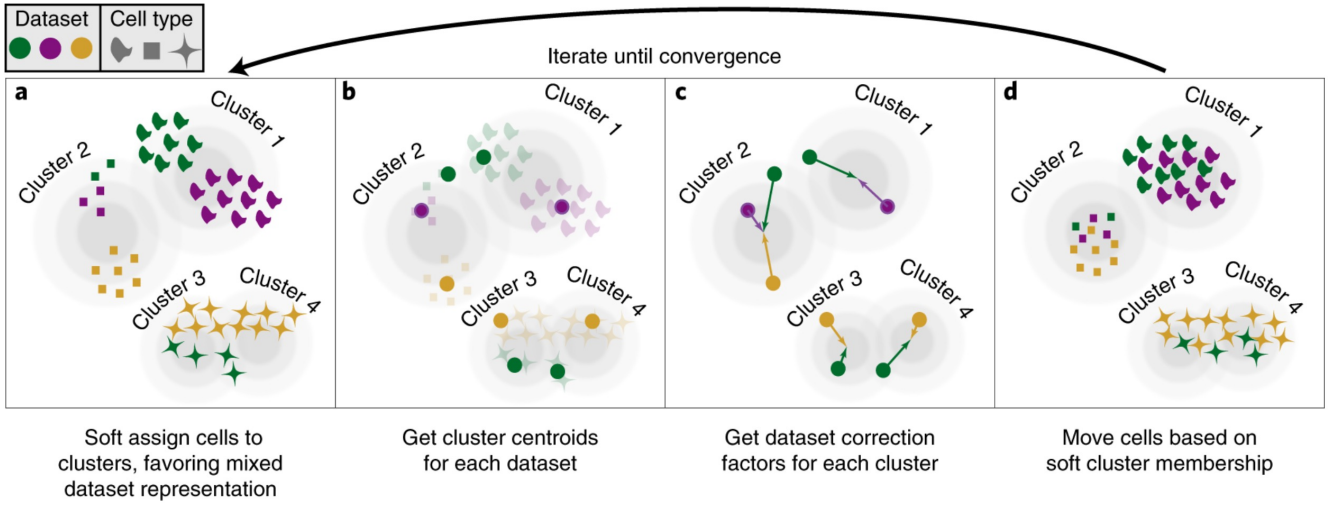


(Stuart et al. 2019, Cell.)

You will try Seurat to integrate scATAC-seq with scRNA-seq data, and do label transfer in an exercise today.

Data integration with Harmony

- An alternative to the Seurat method described above, is Harmony (Korsunsky et al. 2019, Nature Biotech.)
- Assumption: Clusters, representing cell types, should contain cells from many/all batches. → Compute batch correction model based on clustering.



PCA embeds cells into a space with reduced dimensionality. Harmony accepts the cell coordinates in this reduced space and runs an iterative algorithm to adjust for dataset specific effects. **a**, Harmony uses fuzzy clustering to assign each cell to multiple clusters, while a penalty term ensures that the diversity of datasets within each cluster is maximized. **b**, Harmony calculates a global centroid for each cluster, as well as dataset-specific centroids for each cluster. **c**, Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. **d**, Finally, Harmony corrects each cell with a cell-specific factor: a linear combination of dataset correction factors weighted by the cell's soft cluster assignments made in step **a**. Harmony repeats steps **a** to **d** until convergence. The dependence between cluster assignment and dataset diminishes with each round. Datasets are represented with colors, cell types with different shapes.

- Usually works well, much faster than the Seurat method.

Data integration example

- Say that you would have scRNA-seq, scATAC-seq and spatial transcriptomics data.
- If you would successfully integrate these data sets you would have the following information for each cell type:
  - Which genes are expressed (from RNA-seq)
  - Where the open chromatin is, and which TFs might bind there (from ATAC-seq)
  - Where these cells are located (from spatial transcriptomics)

Unsupervised methods

Why unsupervised methods

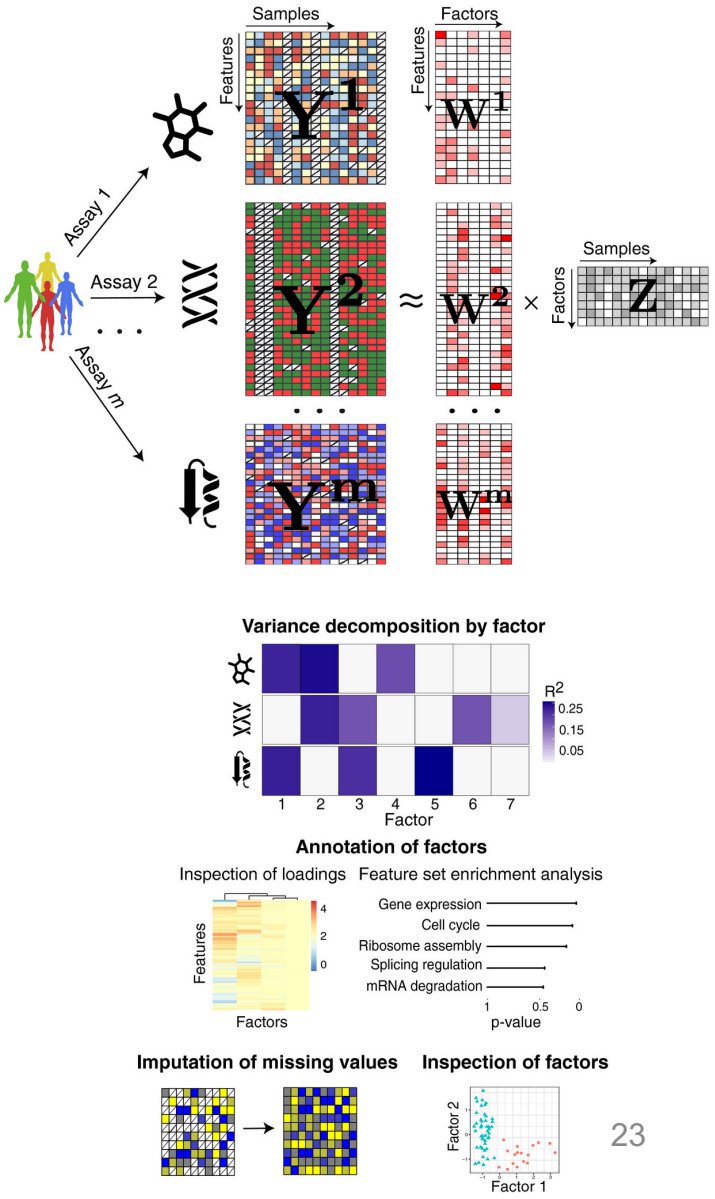
SciLifeLab

- Unsupervised methods are used to find patterns in data, without relying on e.g. a phenotype.
  - Principal Component Analysis (PCA) is an unsupervised method.
- Can be used for
  - Finding important patterns in data (e.g. clusters, trajectories, outliers)
  - Finding features (genes, regions etc.) behind these patterns
  - Imputation
  - Data integration
- Hypothesis-free, good for finding unexpected things!

MOFA I

- Multi-Omics Factor Analysis (MOFA)
- “MOFA can be viewed as a versatile and statistically rigorous generalization of principal component analysis (PCA) to multi-omics data”

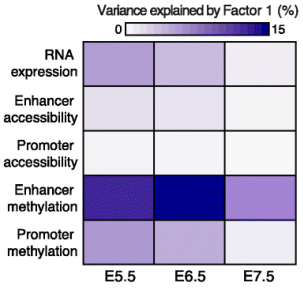
- Compared to PCA:
  - Can handle several types of data at once
    - Each factor can contain all data types
  - Can handle different distributions of data
    - PCA assumes Gaussian distributions
  - Can handle missing values

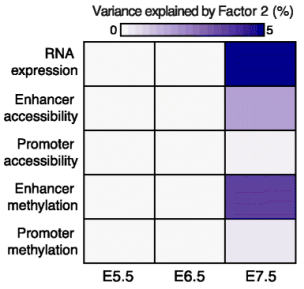


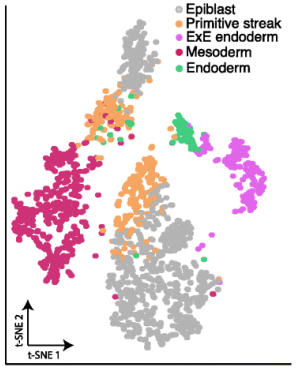
(Argelaguet et al. 2018, Molecular Systems Biology)

MOFA II

- Example: scNMT-seq (single-cell nucleosome, methylation and transcription sequencing)
  - 1828 mouse cells
  - 3 developmental stages
  - RNA expression
  - DNA methylation
  - chromatin accessibility
- Factors capture differentiation, without having access to this information.
- Looking at the weights, we can see how gene expression, chromatin accessibility and DNA methylation interact.





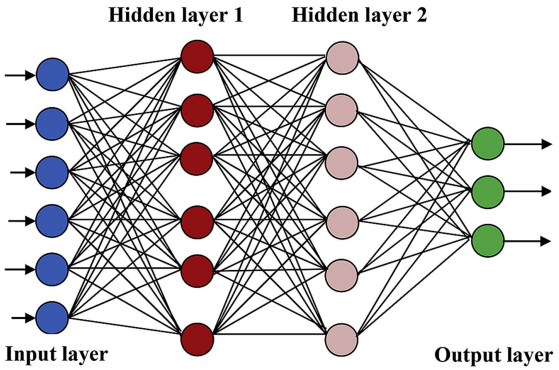


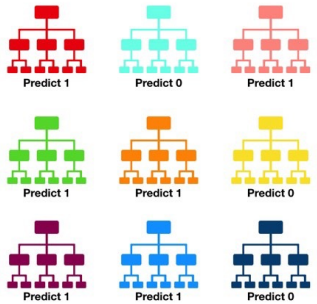
(Argelaguet et al. 2020, Genome Biology)

Supervised methods for data integration

SciLifeLab

- Build models that can combine several data sets to predict an outcome.
  - Example: Use RNA-seq + DNA methylation + H3K27me3 to predict disease outcome
  - Is prediction using all data sets together better than prediction with each individual data set?
  - The model itself can be interesting to investigate: which features are important, which features interact.
- Curse of dimensionality
  - Often many features and few samples
  - Feature selection crucial
- Many different types of methods
  - Random Forest
  - LASSO/Ridge regression etc.
  - PLS/PLS-DA/OPLS
  - Neural networks
- This is a very big topic, outside the scope of this course.





Summary

SciLifeLab

- Many different goals of data integration, and many ways to do it
  - Simple comparisons between data sets (how well do the data sets agree/correlate/overlap?)
  - Integration of single cell data (label transfer etc.)
  - Unsupervised integration (find major patterns in data)
  - Supervised patterns in data (find features informative of a phenotype)
- This is a big subject
  - We have a whole course on data integration if you want to learn more
  - Next occasion: 2023
  - https://uppsala.instructure.com/courses/67276


That's all Folks!