scRNAseq workshop

Yasin Kaymaz

7/27/2018

Single cell RNAseq (scRNAseq) Challenges

- Dropout events / zero inflation
- Cell to cell variation among the same cell types.
- Continuous fuzzy boundaries between transcriptomically distinct cells. No clear discrete boundaries.
- Challenges with integration of different types of single-cell RNAseq data:
  - Batch effect and technical variation.
    - Cell collection method etc.
    - Method used for single-cell: Fluidigm C1, DropSeq, InDrop, 10X etc.
    - Sequencing read features: single-end, paired-end, read length

Goal 1: Which pipeline to use for scRNAseq?

- Alignment-dependent:
  - RSEM-STAR:
    - Aligns reads against a predefined transcriptome sequence.
    - Slow run time.
    - No novel transcript finding.
- Alignment-independent:
  - Salmon, Kallisto:
    - Require reads to "pseudo-map" to transcriptome sequences.
    - Use a transcriptome de Bruijn graph (T-DBG).
    - Increadibly fast!
    - No novel transcript finding.

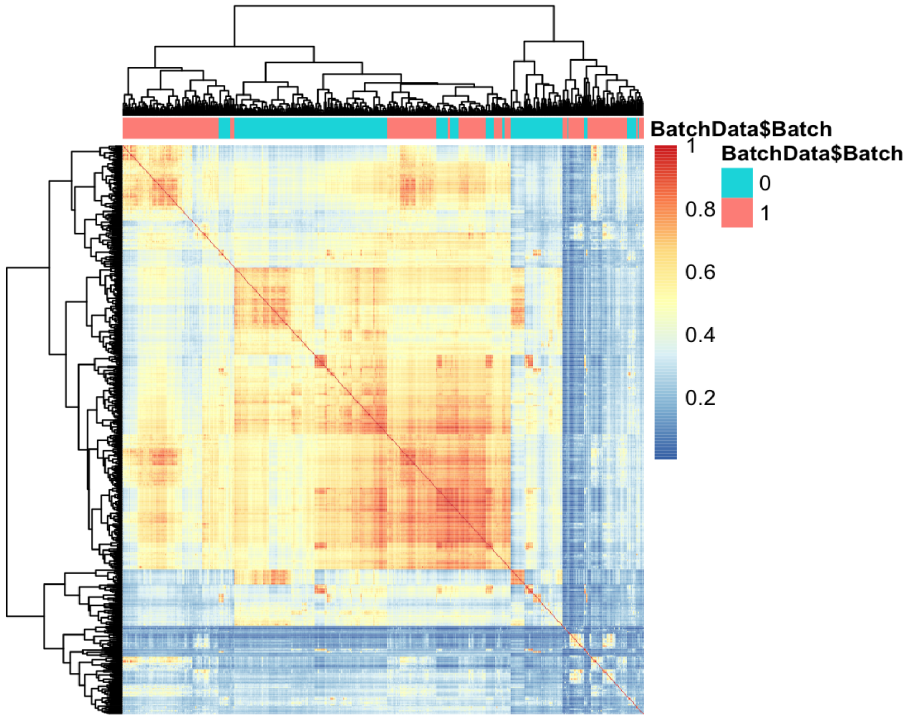
Goal 1: Conclusions:

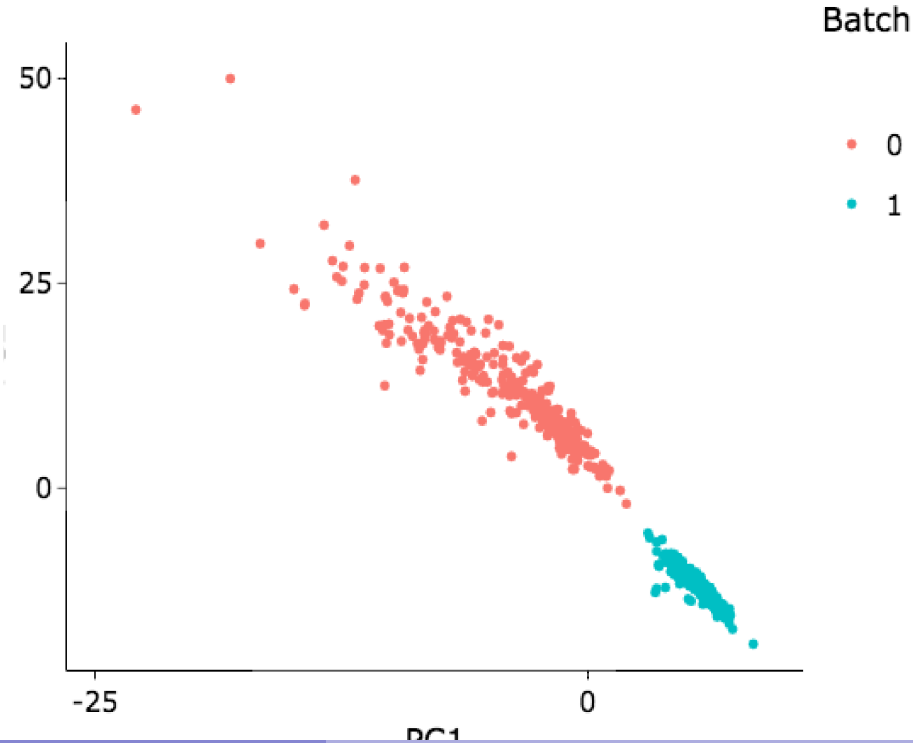
- Rsem, Kallisto, Salmon:
  - All provide similar clusters.
  - Almost same marker genes list.
  - Overall gene Fold changes are highly correlated.
- Salmon and Kallisto are super faster relative to Rsem.
- They recently implemented UMI counting with Kallisto.
- Kallisto also provides an alternative counting scheme:
  - Transcript Compatibility Counts (TCC)
- We feel confident about switching to Kallisto.

Note: Kallisto pipeline is implemented in snakemake.

Goal 2: How to handle datasets from different batches?

- Why do we care?
- Different batches propagate technical variation which can mask true biological signal.





CCA Alignment for Batch integration:

- Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species." Nature Biotechnology. https://doi.org/10.1038/nbt.4096.
- Alignment of shared gene expressions using Dynamic Time Wrapping Algorithm:

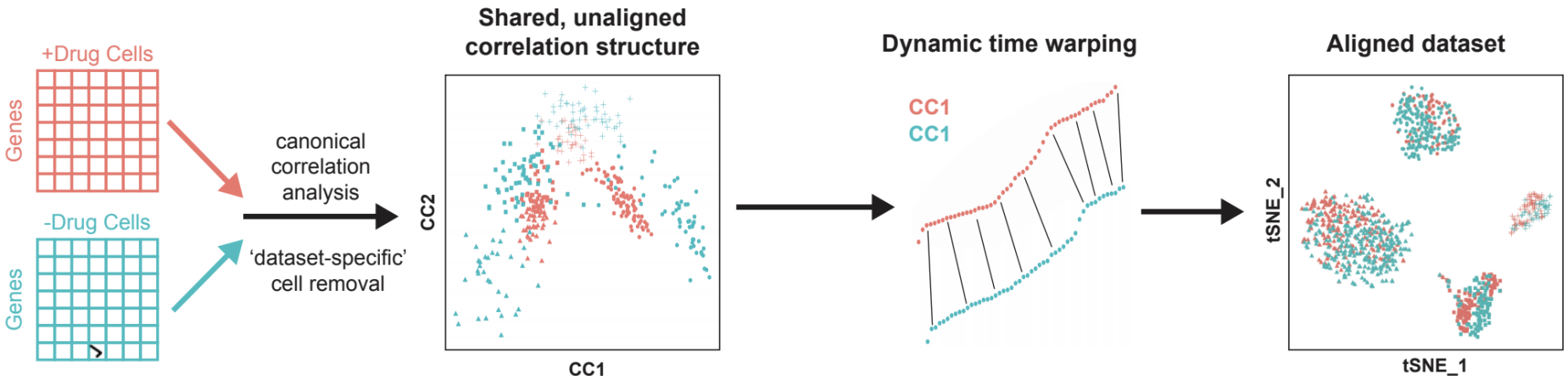


Figure 2: Overview of alignment workflow

CCA Alignment for Batch integration:

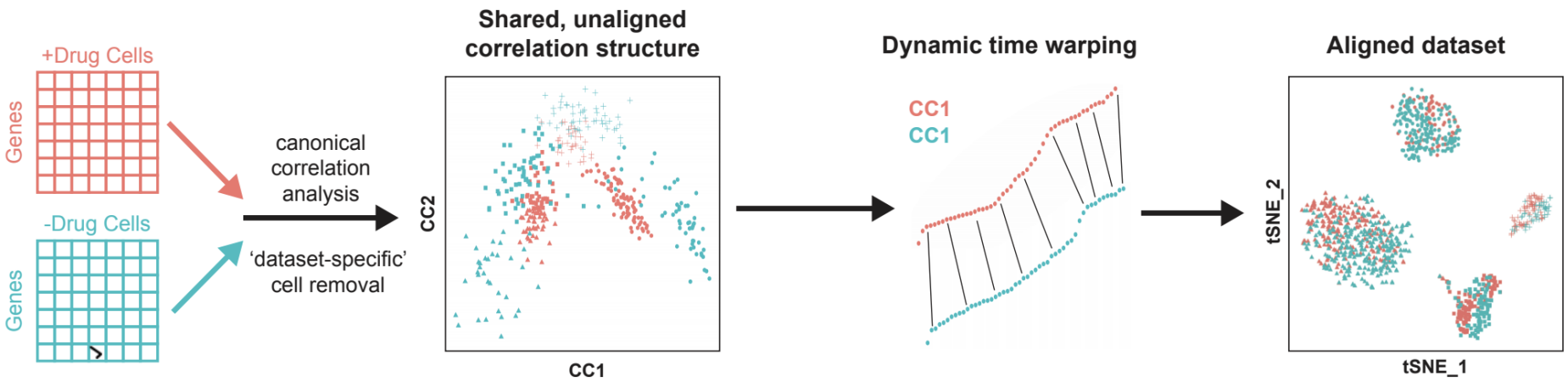


Figure 3: Overview of alignment workflow

1. Select highly variable genes shared by at least two datasets,
2. Identify shared correlation structures (cannonical correlation vectors) across datasets,
3. Align these dimensions using dynamic time wrapping.
4. Use cells embedded into low-dimensional space for clustering.

CCA Alignment for different seq technolojies:

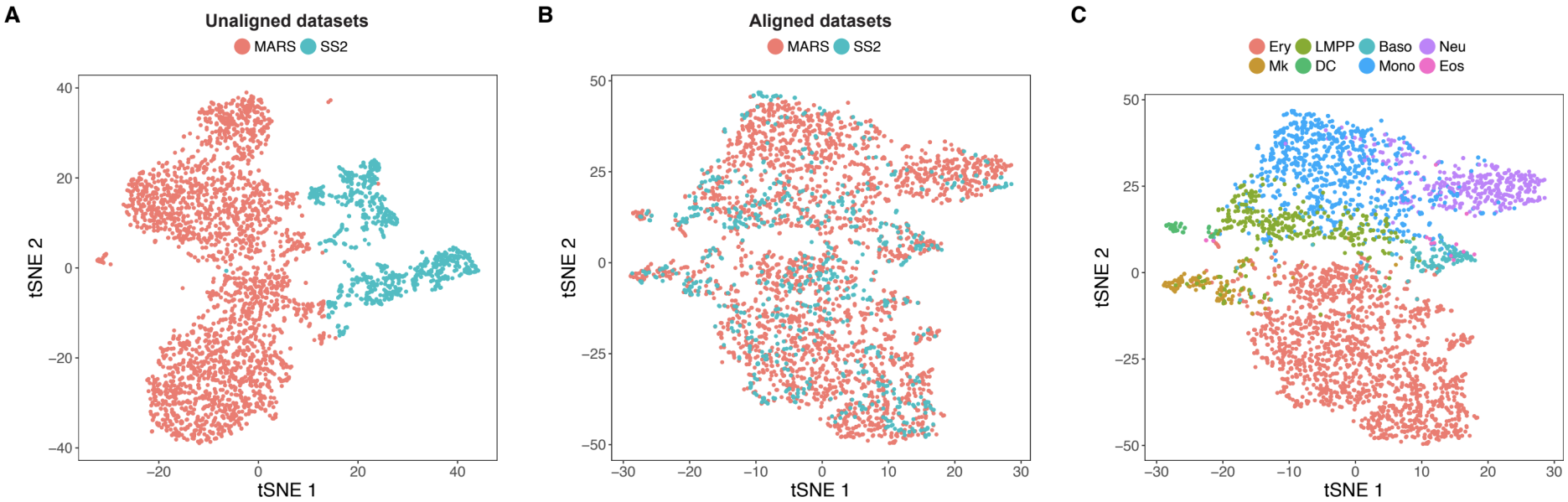


Figure 4: Example 1: Integation of datasets from **different sequencing technologies**

tSNE plots of 3,454 hematopoietic progenitor cells from murine bone marrow sequenced using MARS-seq (2,689) and SMART-Seq2 (765), prior to (A) and post (B-C) alignment. After alignment, cells group together based on shared progenitor type irrespective of sequencing technology.

CCA Alignment for human and mouse datasets:

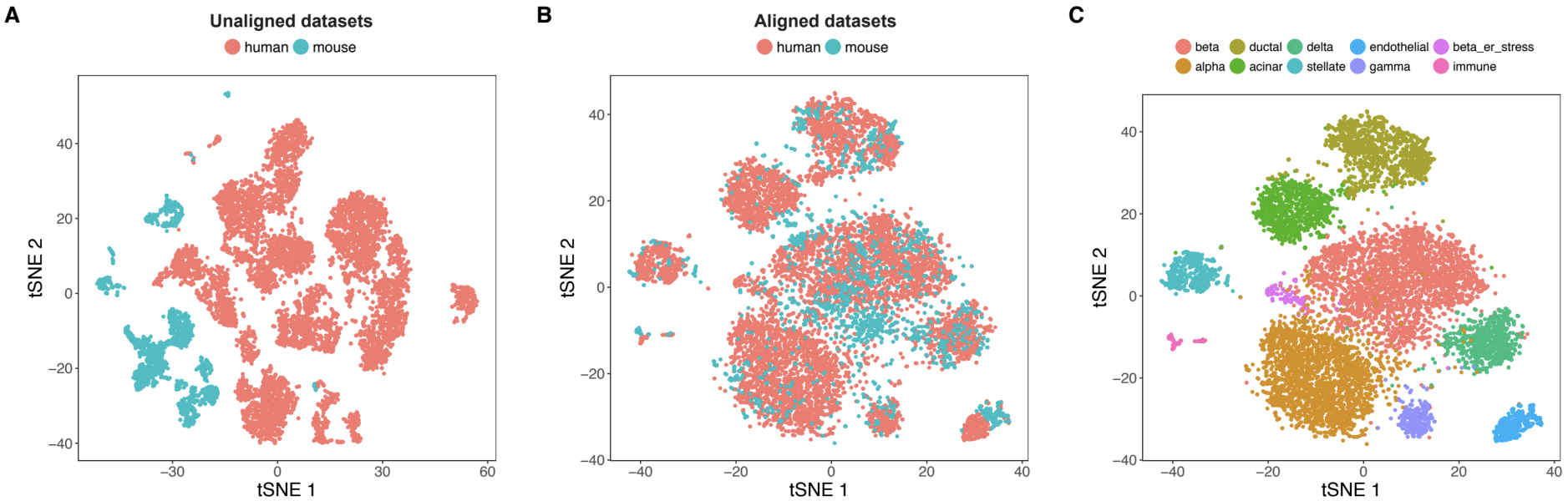


Figure 5: Example 2: Integation of datasets from **human and mouse**

tSNE plots of 10,322 pancreatic islet cells from human (8,533) and mouse (1,789) donors, prior to (A) and post (B) alignment. After alignment, cells group across species based on shared cell type, allowing for a joint clustering (C) to detect 10 cell populations.

Workshop overview:

1. First Part, Basic scRNAseq analysis steps:

- Input data and analysis tool requirements (setup)
  - Datasets and pre-processing steps
- Loading data and initial quality checks
- Filtration, Normalization, and Scaling Expression data
- Finding variable genes and cell subtypes (with tSNE)
- Detecting marker genes of cell clusters

2. Second Part, Working with multiple scRNAseq datasets:

- Loading datasets and QC (with PCA)
- Finding common variable genes between multiple datasets
- Multi-dataset alignment with CCA and Clustering cells (tSNE)
- Before and after comparison (pre/post-Alignment cell clustering)
  - Pros/Cons of this method