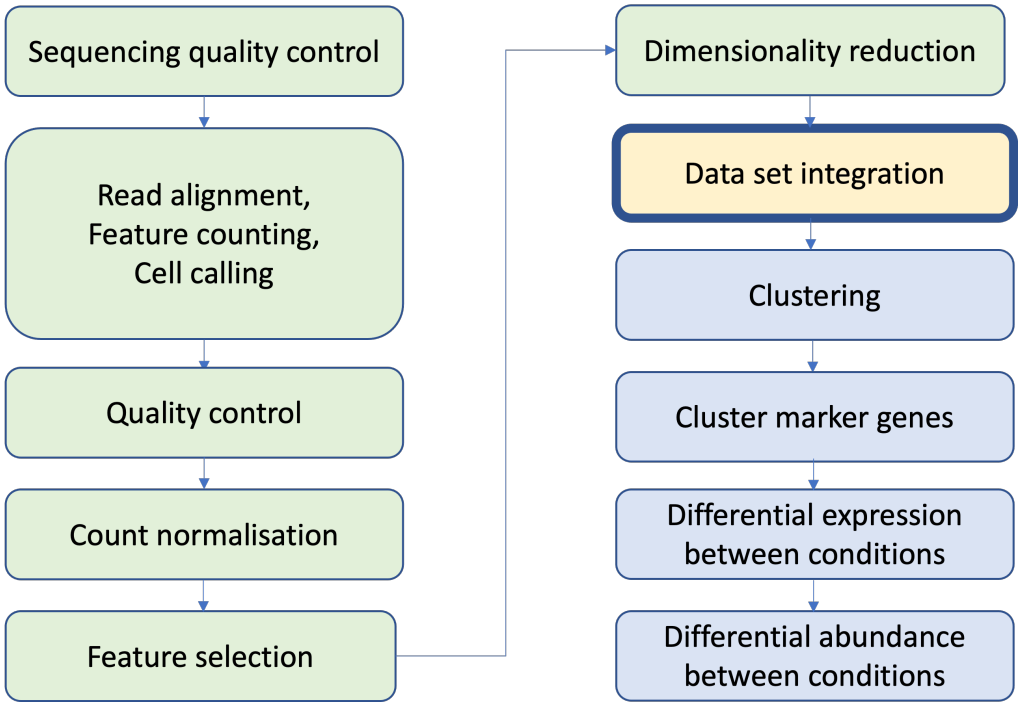
Data Integration and Batch Correction

May 2023

Single Cell RNAseq Analysis Workflow



Why do we need to think about data integration?

- There are generally three reasons for this
  - **Batch effects**:
    - Process samples in batches, different dates, different technicians, different technologies etc
  - **Biological effects**:
    - A study involving male and female subjects with the same disease will often have gender-specific clusters when visualized using t-SNE.
    - Need to integrate to remove the "gender" effect and to identify shared cell types.
  - **Distinct cellular modalities**:
    - For examples for the same study one may profile single cell level transcriptomics or spatial transcriptomics or single cell's immunophenotype
    - Integration is required to to get comprehensive functional understanding of these data sets.

Data Integration Workflow

Formatting our data

A few ways our data can be arranged (software-dependent too)

- one large SCE object containing many samples
- many single-sample SCE objects, QC'd in isolation
- multiple large SCE objects with multiple samples

Important we make sure things match up

- Different bioconductor versions
- Different analysts may have formatted things differently

Cellranger aggr

A useful quick look

10X GENOMICS

Products Research Areas Resources Support Company

Support › Single Cell Gene Expression › Software

SEARCH Q&A NEW CONTACT SUPPORT

Setting Up an Aggregation CSV

Create a CSV file with a header line containing the following columns:

- `sample_id`: Unique identifier for this input GEM well. This will be used for labeling purposes only; it doesn't need to match any previous ID you've assigned to the GEM well.
- `molecule_h5`: Path to the `molecule_info.h5` file produced by `cellranger count`. For example, if you processed your GEM well by calling `cellranger count --id=ID` in some directory `/DIR`, this path would be `/DIR/ID/outs/molecule_info.h5`.

You can either make the CSV file in a text editor, or create it in Excel and export to CSV. Continuing the example from the previous section, your Excel spreadsheet would look like this:

	A	B
1	sample_id	molecule_h5
2	LV123	/opt/runs/LV123/outs/molecule_info.h5
3	LB456	/opt/runs/LB456/outs/molecule_info.h5
4	LP789	/opt/runs/LP789/outs/molecule_info.h5

When you save it as a CSV, the result would look like this:

```
sample_id,molecule_h5
LV123,/opt/runs/LV123/outs/molecule_info.h5
LB456,/opt/runs/LB456/outs/molecule_info.h5
LP789,/opt/runs/LP789/outs/molecule_info.h5
```

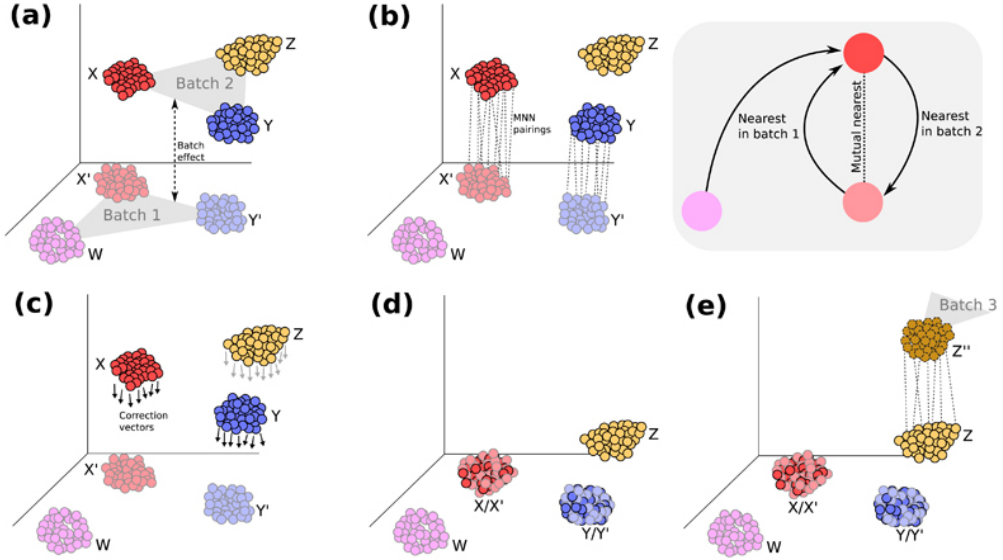
Cell Ranger v6.0 (latest)

Checking for batch effects

Batch Corrections

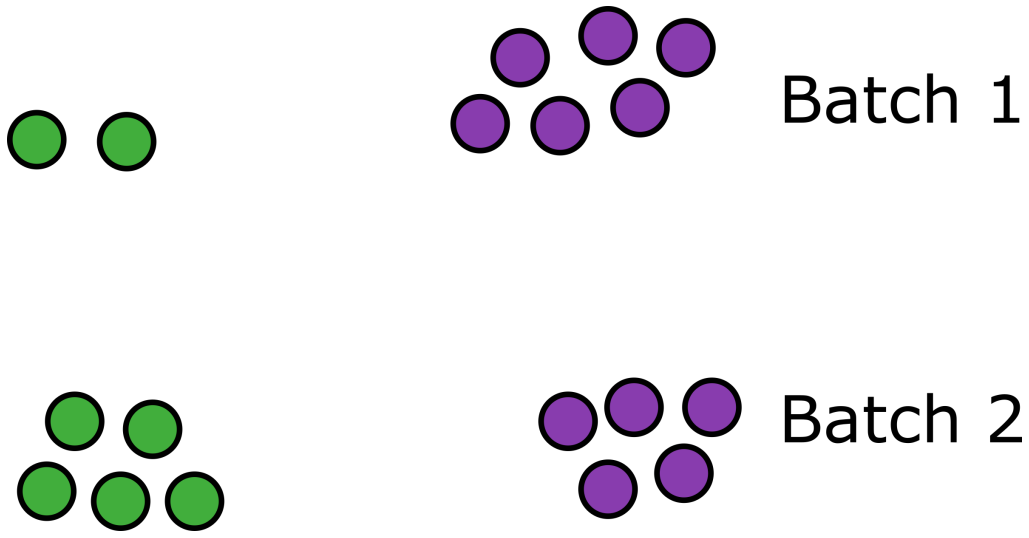
- Gaussian/Linear Regression - removeBatchEffect (limma), comBat (sva), rescaleBatches or regressBatches (batchelor)
- Mutual Nearest Neighbours (MNN) correction - Haghverdi et al 2018
  - mnnCorrect (batchelor)
  - FastMNN (batchelor)
- And many more!
  - Different methods may have strenghts and weaknesses
  - Benchmark studies can be used as a reference to choose suitable method

FastMNN (Haghverdi et al 2018)

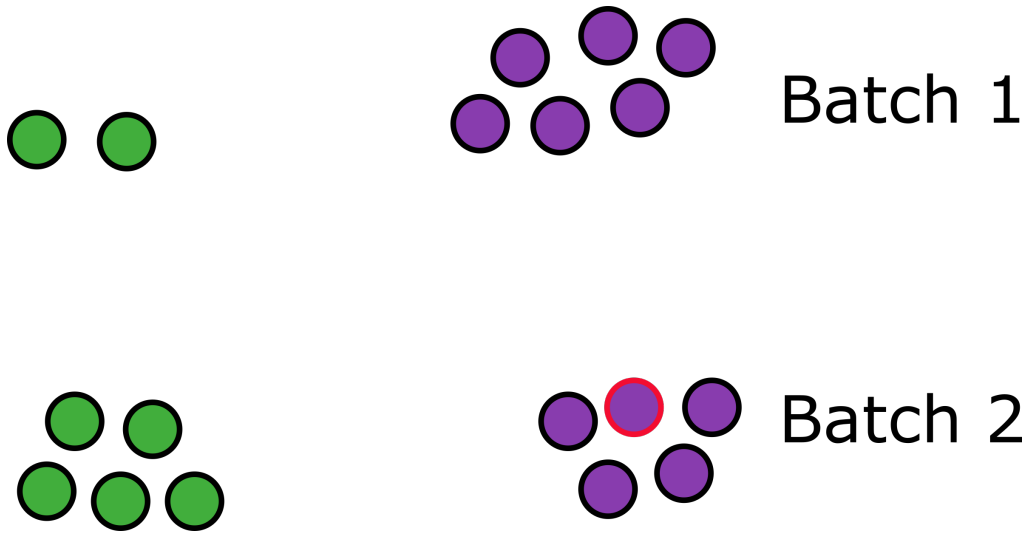


1. Perform a multi-sample PCA on the (cosine-)normalized expression values to reduce dimensionality.
2. Identify MNN pairs in the low-dimensional space between a reference batch and a target batch.
3. Remove variation along the average batch vector in both reference and target batches.
4. Correct the cells in the target batch towards the reference, using locally weighted correction vectors.
5. Merge the corrected target batch with the reference, and repeat with the next target batch.

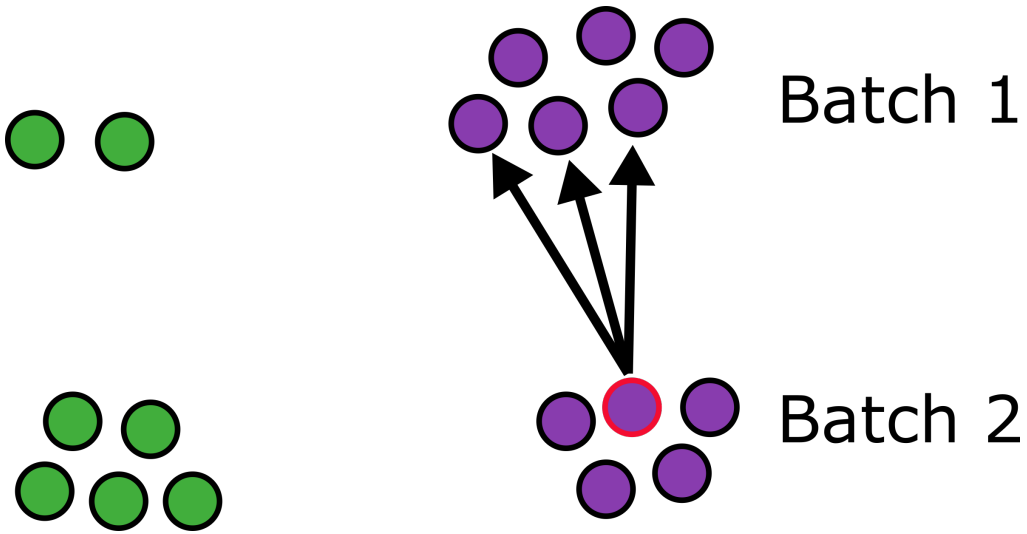
FastMNN (Haghverdi et al 2018)



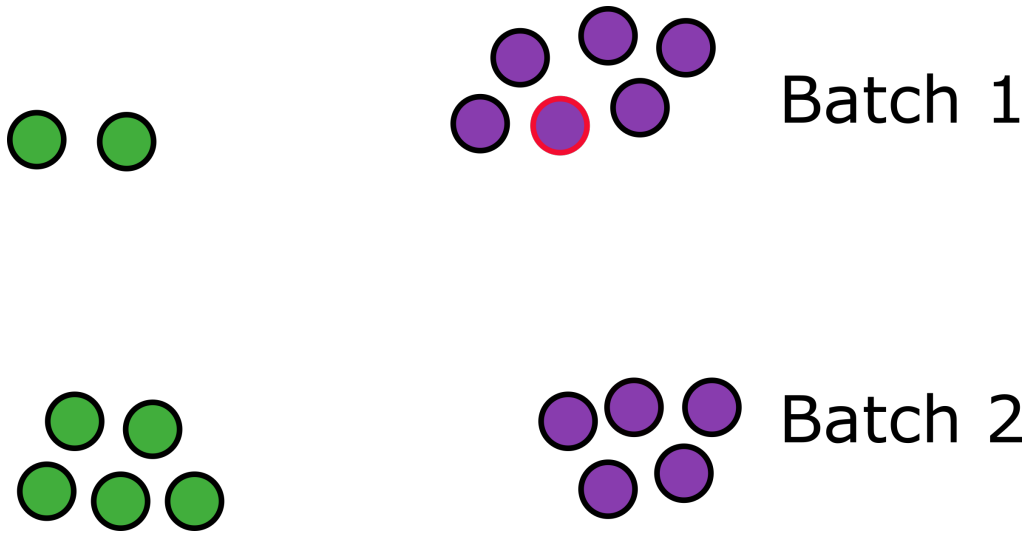
FastMNN (Haghverdi et al 2018)



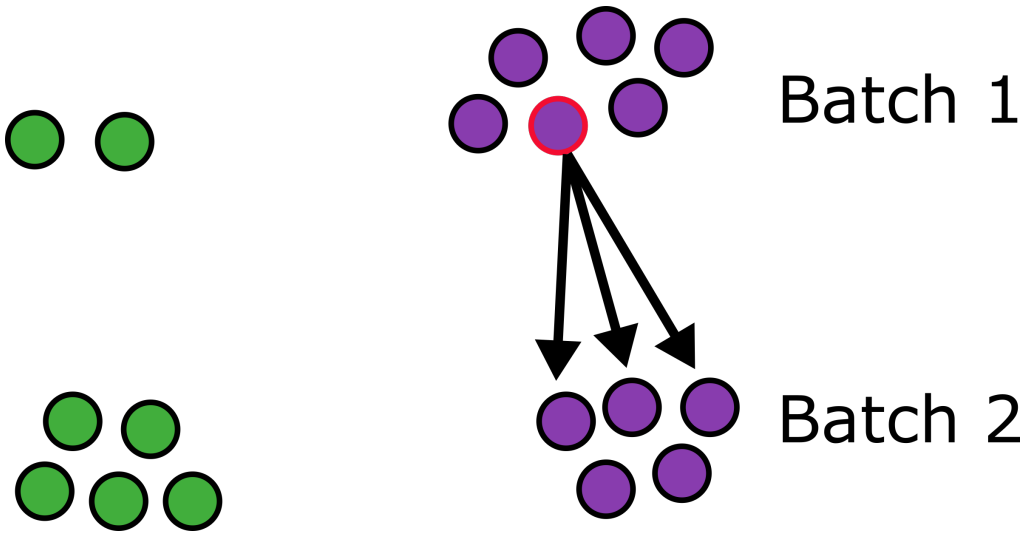
FastMNN (Haghverdi et al 2018)



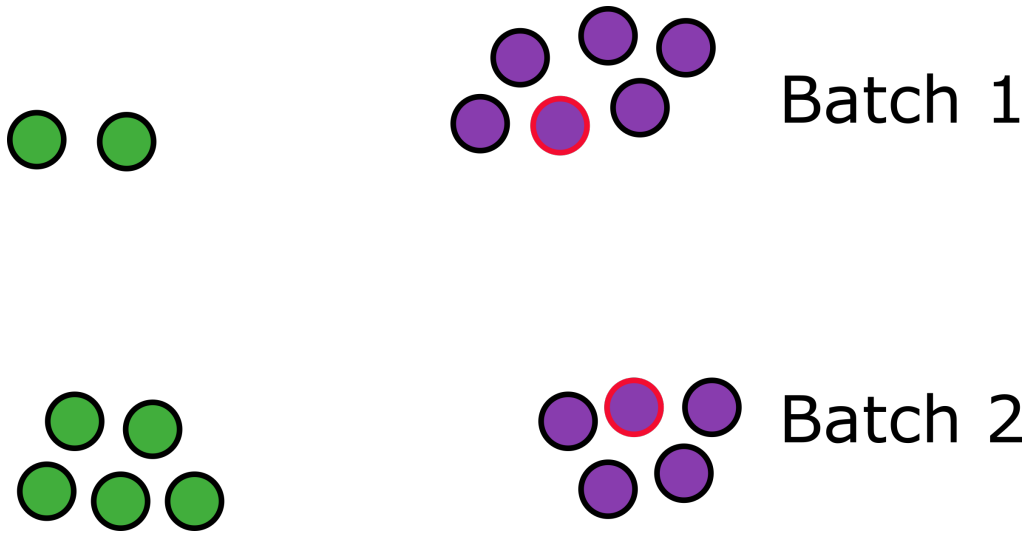
FastMNN (Haghverdi et al 2018)



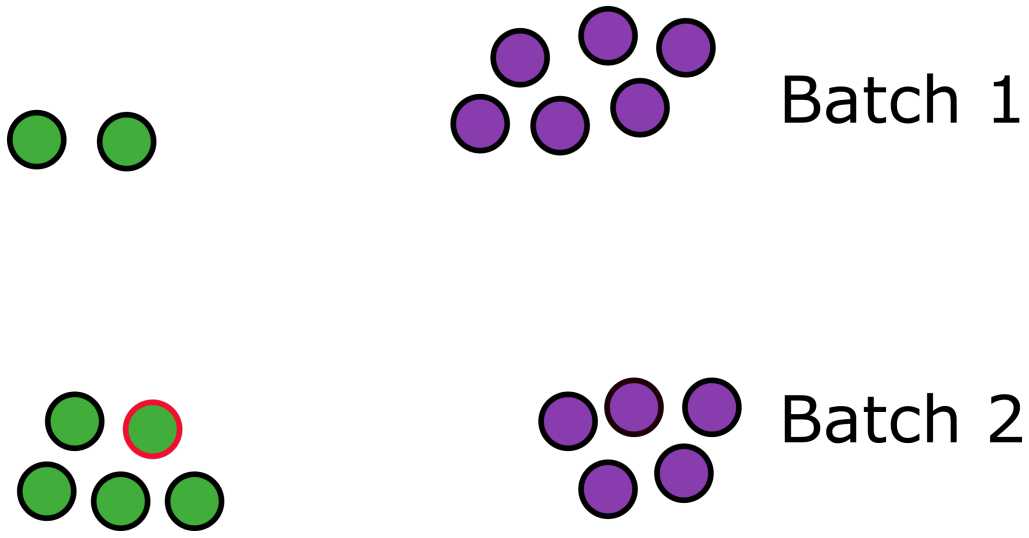
FastMNN (Haghverdi et al 2018)



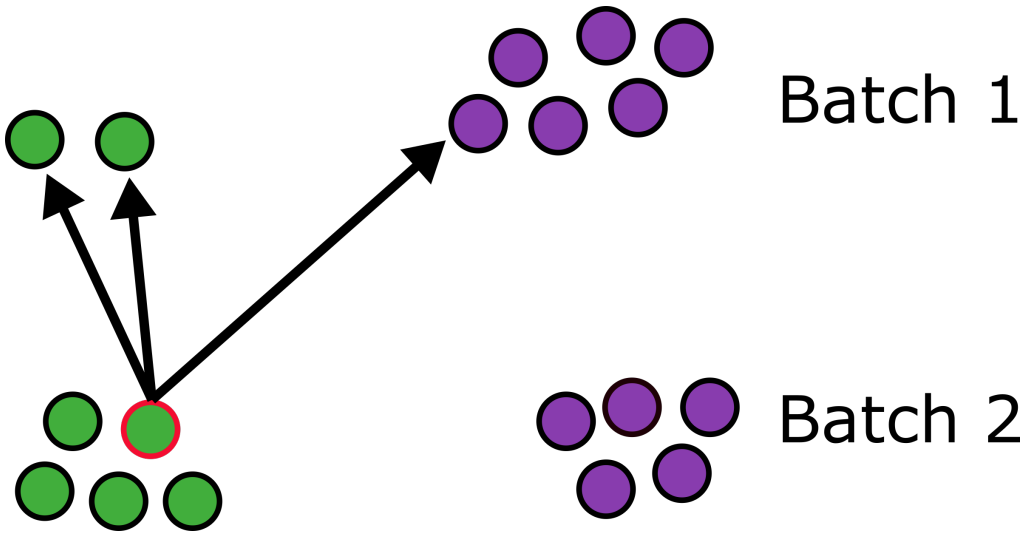
FastMNN (Haghverdi et al 2018)



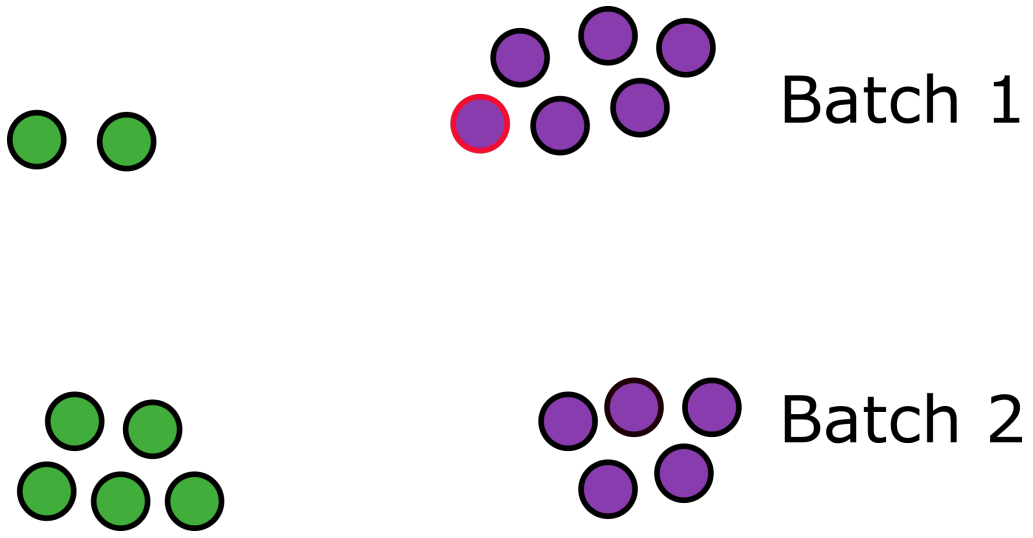
FastMNN (Haghverdi et al 2018)



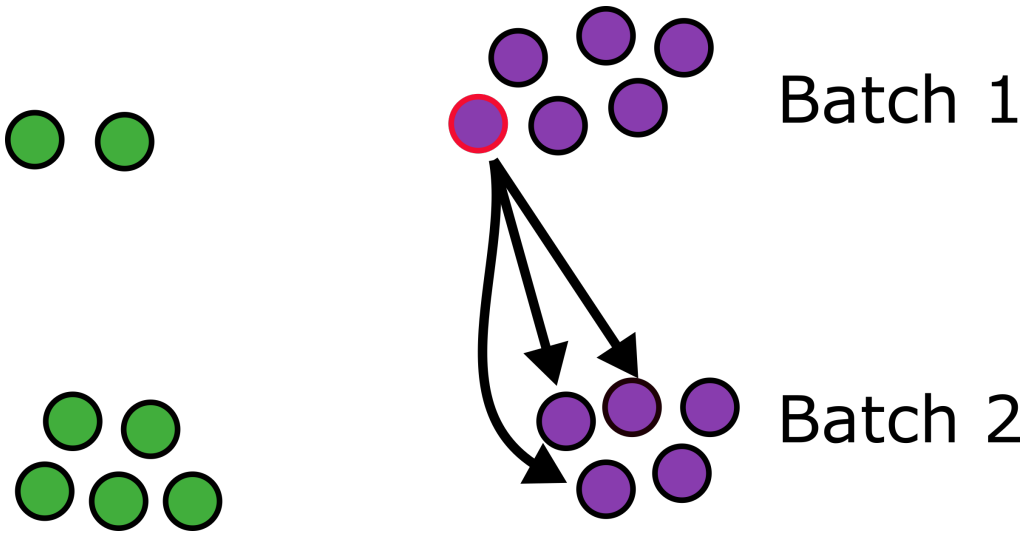
FastMNN (Haghverdi et al 2018)



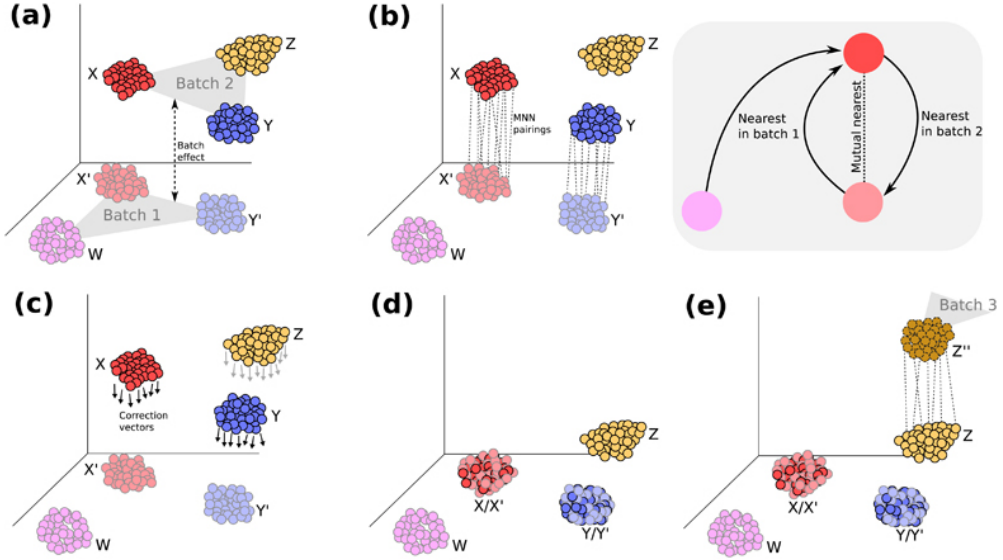
FastMNN (Haghverdi et al 2018)



FastMNN (Haghverdi et al 2018)

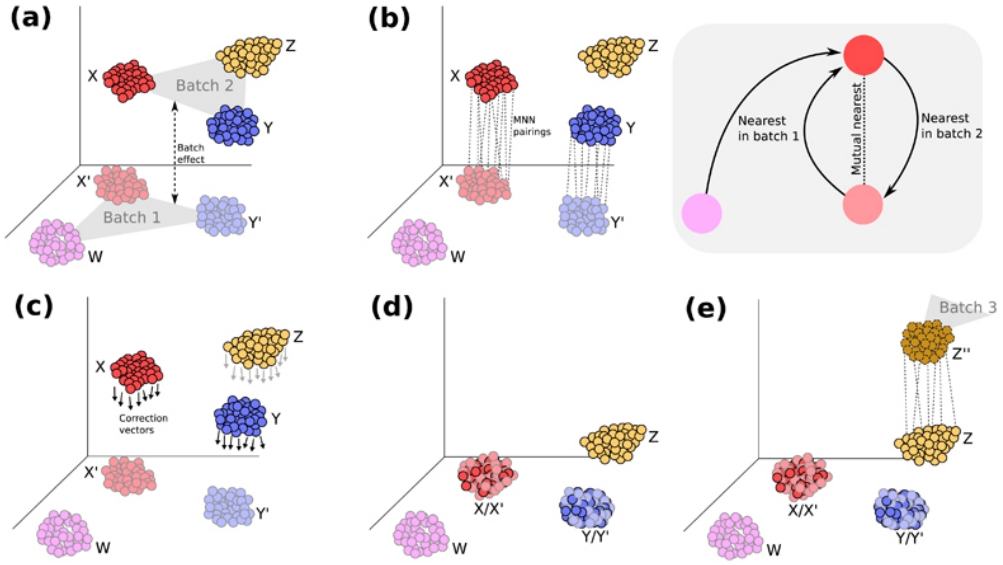


FastMNN (Haghverdi et al 2018)



1. Perform a multi-sample PCA on the (cosine-)normalized expression values to reduce dimensionality.
2. Identify MNN pairs in the low-dimensional space between a reference batch and a target batch.
3. Remove variation along the average batch vector in both reference and target batches.
4. Correct the cells in the target batch towards the reference, using locally weighted correction vectors.
5. Merge the corrected target batch with the reference, and repeat with the next target batch.

FastMNN (Haghverdi et al 2018)



Assumptions (quoted from the paper):

1. There is at least one cell population that is present in both batches,
2. the batch effect is almost orthogonal [i.e. uncorrelated] to the biological subspace, and
3. the batch-effect variation is much smaller than the biological-effect variation between different cell types

Checking our correction has worked

Checking our correction has worked

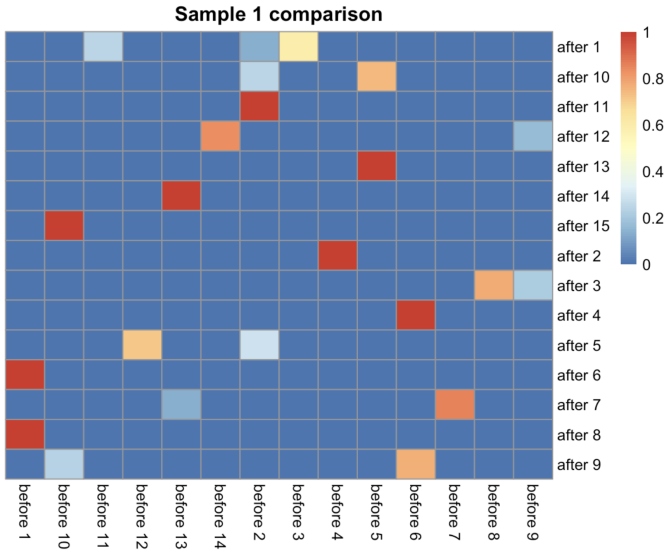
We can look at the 'mixing' between batches and calculate the variance in the log-normalized cell abundances across batches for each cluster.

Clusters are ranked by variance for manual inspection.

If variance is too high it could indicate there isn't sufficient correction.

```
##        Batch
## Cluster ETV6-RUNX1_1 ETV6-RUNX1_2 ETV6-RUNX1_3 ETV6-RUNX1_4 HHD_1 HHD_2 PBMMC_1
##       7           341          355          195          202   253   393      68
##       5             0            0            1            0     1     0       1
##       15            4            9          170           27    21     2      62
```

Checking our correction has worked



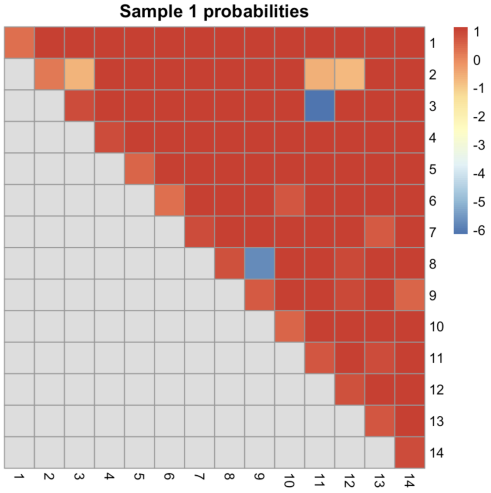
- Investigating which clusters from before correction are nested inside the clusters after correction can help us decide if our correction has worked.
- Did the whole before cluster from one sample go into an after cluster or was it broken apart?
- Perfect nesting would be indicated by one orange/red block in each row
- But do we want that?

Checking our correction hasn't over worked

- If you use fastMNN in the absence of a batch effect, it may not work correctly
- It is possible to remove genuine biological heterogeneity
- fastMNN can be instructed to skip the batch correction if the batch effect is below a threshold. You can use the effect sizes it calculates to do this.
- In reality the absence of any batch effect would warrant further investigation.

Checking our correction hasn't over worked

- One way to measure if we have retained heterogeneity is to look at the agreement between clusters before and after correction
- Adjusted Rand Index
- HIGH = GOOD (eg. 0.8 = within batch variation is retained)



- ARI can also be broken down into per-cluster ratios

Checking our correction hasn't over worked

- There is also an MNN specific metric we can calculate called 'lost variance'
- How much variance within each batch has been removed by the correction
- Ideal < 0.1 or 10%
- Higher levels indicate artificial smoothing of data

Using the corrected values

The value in batch correction is that it enables you to see population heterogeneity within clusters/celltypes across batches.

- Also increases the number of cells you have

However the corrected values should not be used for gene based analysis eg. DE/marker detection.

- fastMNN doesn't preserve the magnitude or direction of per-gene expression and may have introduced artificial agreement between batches on the gene level.