Supplementary material: Correcting batch effects in single-cell RNA sequencing data by matching mutual nearest neighbours

Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, John C. Marioni

1 Cosine normalization and cosine distance

The Euclidean distance between cells x and y following cosine normalisation ($D^2(x,y)$), is equivalent to using the cosine distances between the cells.

$$\begin{aligned} D^2(x,y) &= \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} - \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||}\right)^2 + \left(\frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 - 2\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||}\cdot\frac{\mathbf{Y_y}}{||\mathbf{Y_y}||} \\ &= 2\left(1 - \frac{\mathbf{Y_x}\cdot\mathbf{Y_y}}{||\mathbf{Y_x}||||\mathbf{Y_x}||}\right) = 2\cdot\text{cosine.distance}(x,y) \end{aligned} \tag{1}$$

2 MNN identifies biologically matching cell populations: proof

Here, we show if a cell population (same biological cell type) is measured in two different batches X and Y (assumption i in the main text), they will be identified as nearest neighbours of each other. A required assumption (assumption ii in the main text) is orthogonality of the biological subspace B to the batch specific effects W_X and W_Y. This is a weak assumption as B and W_X and W_Y are usually of low intrinsic dimensionality in the high dimensional space of gene expression.

More specifically the $G * N_b$ dimensional batch data can be explained as a linear combination of a biological B and a batch specific W_b signal, where G is the number of genes measured and N_b is the number of cells in batch b.

$$X = B\beta + W_X\alpha \tag{2}$$

$$Y = B\gamma + W_Y\zeta \tag{3}$$

In the above equations, B is $G * K$, with K being the intrinsic dimensionality of the biological signal. β and γ are $K * N_X$ and $K * N_Y$ matrices respectively. W_X is $G * J_X$, W_Y is $G * J_Y$ and α and ζ are $J_X * N_X$ and $J_Y * N_Y$ respectively, J_b being the intrinsic dimensionality of the batch effect in batch b.

The distance between two cells $x \in X, \quad y \in Y$ is calculated as:

$$\begin{aligned} D^2(x,y) &= (\quad B(\beta_x - \gamma_y) + (W_X\alpha_x - W_Y\zeta_y) \quad)^T(\quad B(\beta_x - \gamma_y) + (W_X\alpha_x - W_Y\zeta_y) \quad) \\ &= (\beta_x - \gamma_y)^T B^T B(\beta_x - \gamma_y) + (W_X\alpha_x - W_Y\zeta_y)^T(W_X\alpha_x - W_Y\zeta_y) \\ &\quad + (\beta_x - \gamma_y)^T B^T(W_X\alpha_x - W_Y\zeta_y) + (W_X\alpha_x - W_Y\zeta_y)^T B(\beta_x - \gamma_y) \\ &= (\beta_x - \gamma_y)^T B^T B(\beta_x - \gamma_y) + (W_X\alpha_x - W_Y\zeta_y)^T(W_X\alpha_x - W_Y\zeta_y) + 0 + 0 \end{aligned} \tag{4}$$

In the last line we have used the assumption of orthogonality of the biological subspace B to the noise or batch specific signals W_X and W_Y. In other words, $B^T \cdot W_X = 0$ and $B^T \cdot W_Y = 0$. We now search

for the shortest distance (i.e. nearest neighbours) to cell $x \in X$ among all cells in batch Y by setting the derivative of the distance (with respect to y) to zero:

$$\frac{dD_x^2}{dy} = \frac{\partial D_x^2}{\partial \gamma_y} + \frac{\partial D_x^2}{\partial \zeta_y} = 0 \quad \Rightarrow \quad B^T B(\beta_x - \gamma_y) + \frac{\partial D_x^2}{\partial \zeta_y} = 0 \tag{5}$$

, where D_x^2 denotes the distance to the cell $x \in X$. Equation 5 will be satisfied if:

$$B^T B(\beta_x - \gamma_y) = -\frac{\partial D_x^2}{\partial \zeta_y} \tag{6}$$

We note that the batch effect in one batch will typically vary not strongly among neighbouring cell types, whereas variability due to the biological signal can be quite large (assumption iii in the main text). That is:

$$\frac{\partial D_x^2}{\partial \zeta_y} \approx 0 \quad \text{for any} \quad y \tag{7}$$

Putting this into equation 6, we conclude:

$$\beta_x - \gamma_{y^*} \approx 0 \tag{8}$$

is the necessary condition for equation 5. Thus, the cell in batch Y at minimum distance to x is y^* with almost similar biological coefficients γ_{y^*} to that of x (i.e. β_x). Similarly, one can show $\beta_{x^*} - \gamma_{y*} = 0$ is the solution when we fix cell $y* \in Y$ and consider the derivative of distance with respect to x instead of y in equation 5.

The above proves that biologically similar cells are exclusively identified as mutual nearest neighbours. For cells that do not have any counterpart in the other batch, the two directional condition in equation 6 for $(\beta_{x*} - \gamma_{y*}) \neq 0$ can only be satisfied at the facing boundaries of two cell populations where there is a discontinuity in $\frac{\partial D_{x*}^2}{\partial \zeta_y}$ and $\frac{\partial D_{y*}^2}{\partial \alpha_x}$. Although it is possible that some mutual nearest neighbours are identified at such regions, these pairs will be restricted to a small region of data corners, thus consist of few cells. This is in contrast to truly biologically related pairs where the cells are distributed across a wide area, sometimes a whole cell population.

3 Algorithm

Algorithm 1 The MNN batch correction algorithm

INPUT: n batches $B_1, ..., B_n$ of correspondingly $N_1, ..., N_n$ cells and G_I inquiry genes, and a set of G_{HVG} highly variable genes.

OUTPUT: Batch corrected and integrated data C of G_I genes and $N_1 + ... + N_n$ cells

for each batch B_i **do**
- Cosine normalise the expression data of the highly variable genes (HVG).

end for

- Define the first batch as the reference data $C \leftarrow B_1$.

for $i = 2 : n$ **do**

- Using k nearestes neighbouring cells search between the batches C and B_i, find corresponding MNN pairs l and m in the cosine normalised HVG genes set. Then calculate the corresponding batch vectors $\vec{v}_I^{ml}$ and $\vec{v}_{HVG}^{ml}$ in the inquiry and HVG genes sets.

for each cell x in B_i **do**

- Calculate the Gaussian kernel weights $W_{HVG}(x, m)$ to all cells in the MNN set of B_i (as found in the previous step).
- Calculate the batch vectors in the inquiry and cosine normalised HVG genes space for x as : $\vec{u}_I(x) = \frac{\sum_m \vec{v}_I^{ml} W_{HVG}(x,m)}{\sum_m W_{HVG}(x,m)}$ and $\vec{u}_{HVG}(x) = \frac{\sum_m \vec{v}_{HVG}^{ml} W_{HVG}(x,m)}{\sum_m W_{HVG}(x,m)}$.
- Calculate the biological subspace for each MNN set as the d first SVDs of centered expressions of cells $l \in C$ and $m \in B_i$ (for the inquiry set of genes and the HVG genes).
- Remove the biological components ($\vec{biol}$) from the batch vectors $\vec{u}_I(x) \leftarrow \vec{u}_I(x) - \vec{biol}_I$ and $\vec{u}_{HVG}(x) \leftarrow \vec{u}_{HVG}(x) - \vec{biol}_{HVG}$
- Correct for the batch effect in inquiry genes and cosine normalised HVG genes sets for $x \in B_i$ by $\vec{x}_I \leftarrow \vec{x}_I - \vec{u}_I(x)$ an $\vec{x}_{HVG} \leftarrow \vec{x}_{HVG} - \vec{u}_{HVG}(x)$.

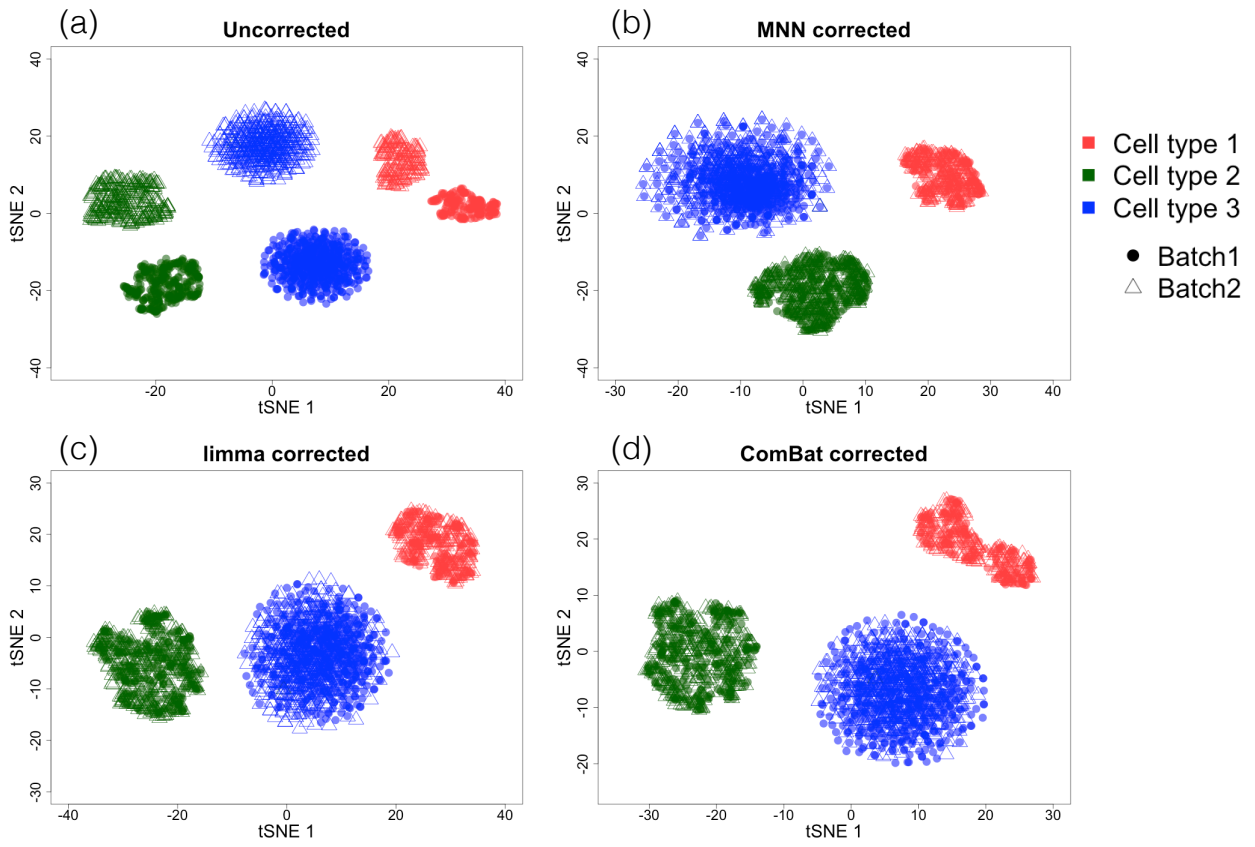
end for

- Append the corrected data B_i' to C.

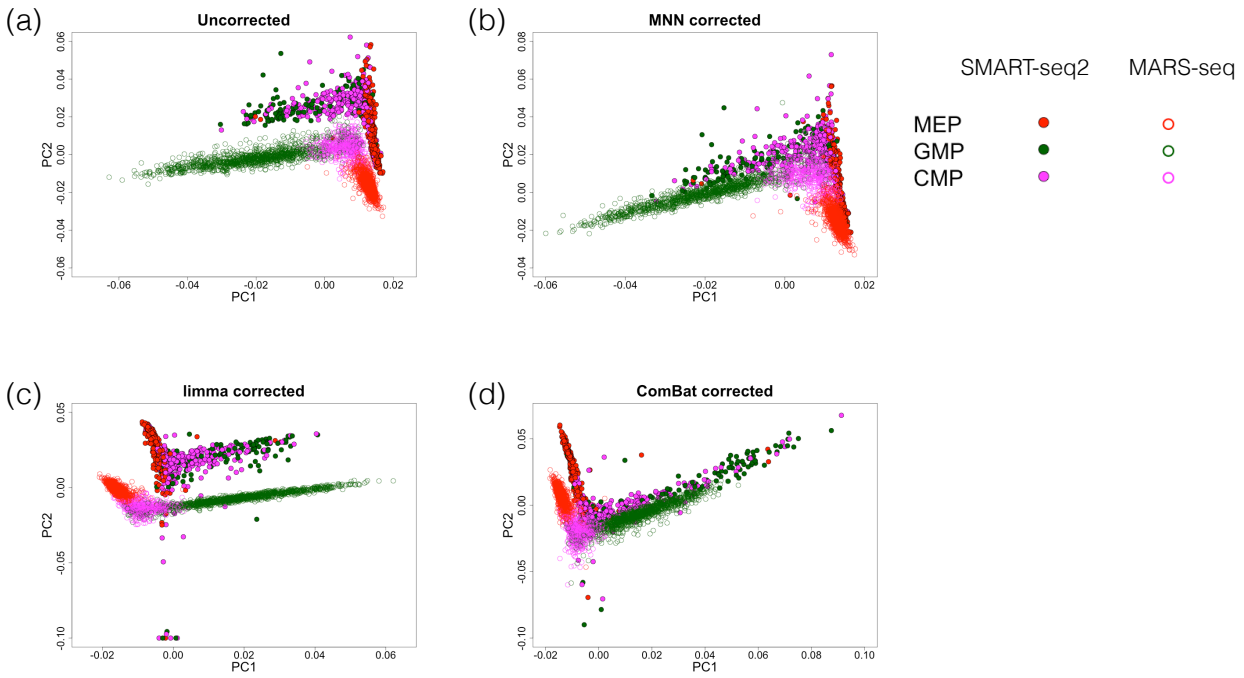
end for

The default setting of the parameters is as follows:

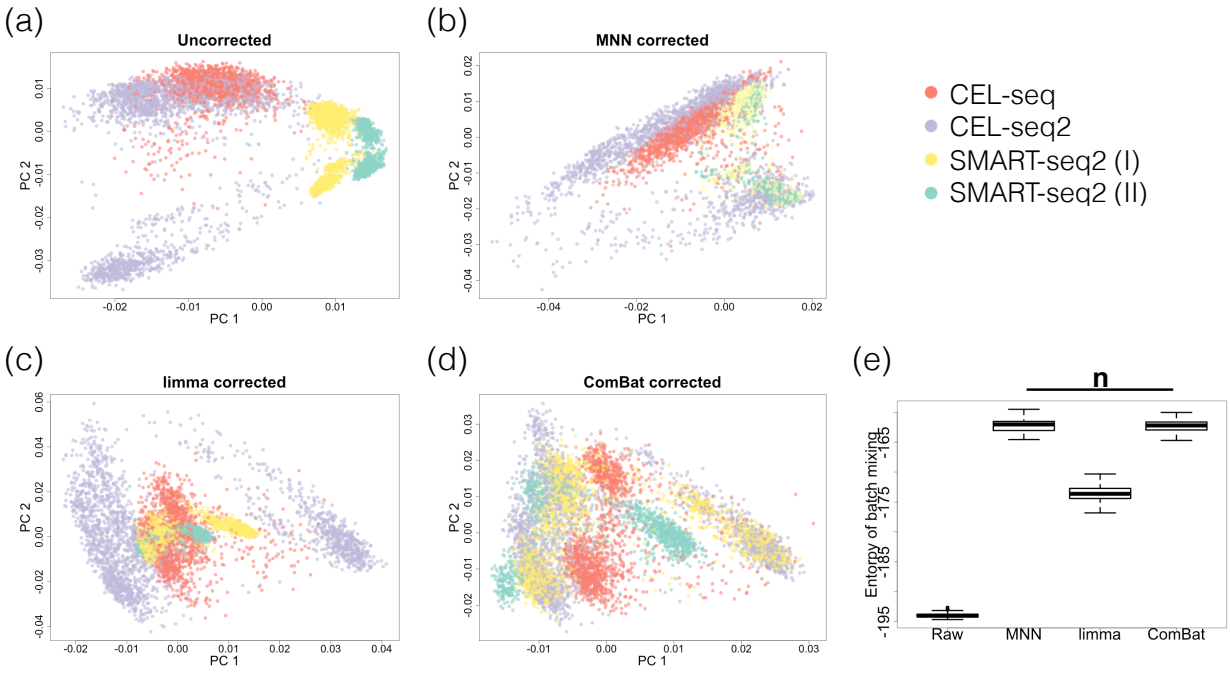
- The default number of nearest neighbours search for identification of MNN pairs is $k = 20$. This number might need to be changed depending on the data size, but should not be too big such that every cell in one batch is among the nearest neighbours of the every cell in the other batch. A too small k (e.g. $k = 1$) will also not be suitable as it would imply a too sensitive MNN identification criteria on the noisy expression data, such that no MNNs can eventually be identified.
- We choose the default Gaussian kernel width $\sigma^2 = 0.1$ in the cosine normalized space, where the L^2 norm of any expression vector is equal to 1.
- The default number of SVDs assumed to capture the biological subspaces is $d = 2$.



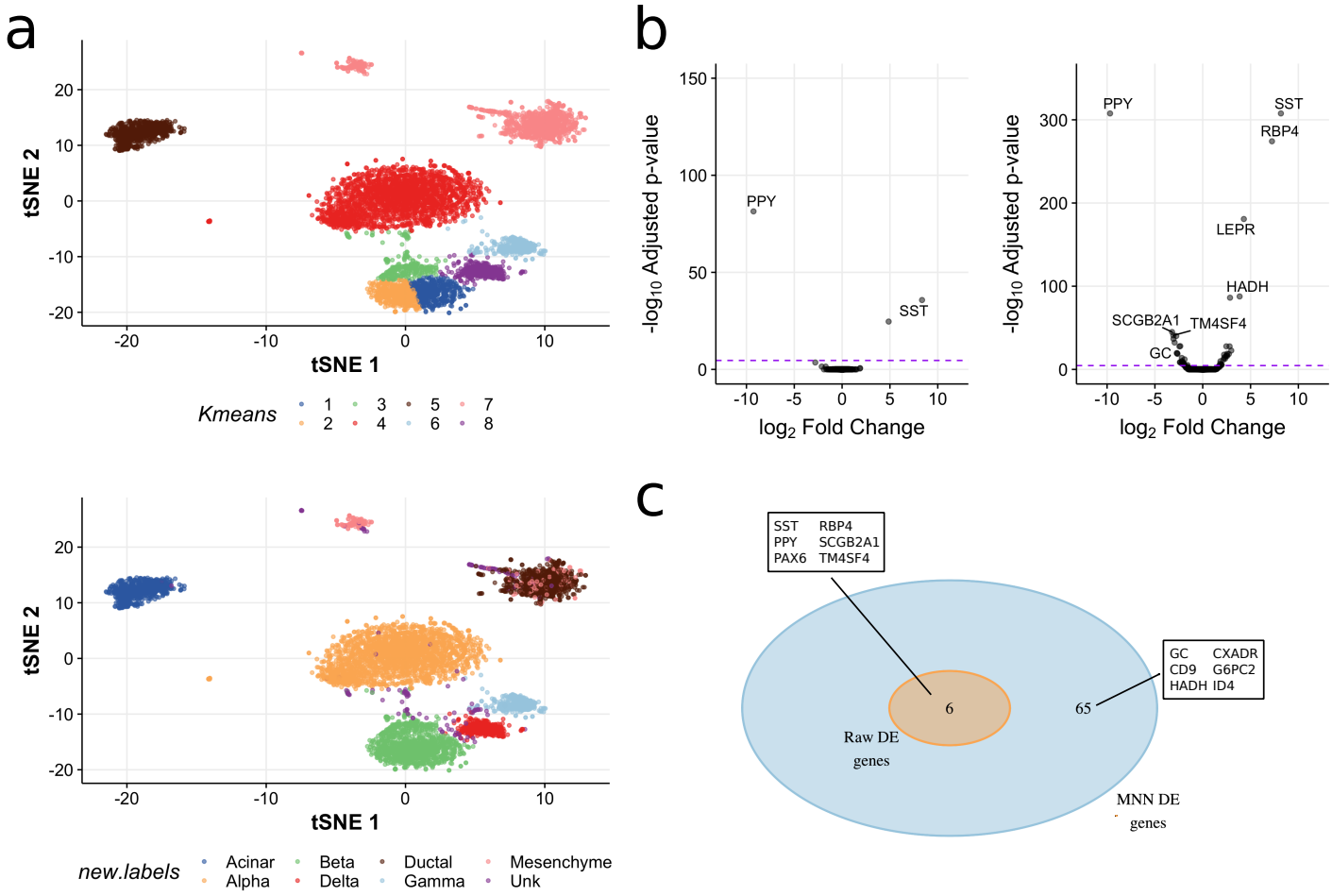
Supplementary Figure 1: t-SNE plots of (a) the raw (uncorrected) simulated data, and the simulation data corrected by (b) MNN ,(c) limma and (d) ComBat. The filled circles and open triangles represent cells from the first and second batch respectively. The three different cell types are shown by different colors. While there is a split between cells of the same cell type in the uncorrected data, all batch correction methods remove the batch effect successfully for this simple example and result in clustering of data according to the original simulated cell types. The data were simulated to have identical cell type compositions (0.2/0.3/0.5) between the two batches.



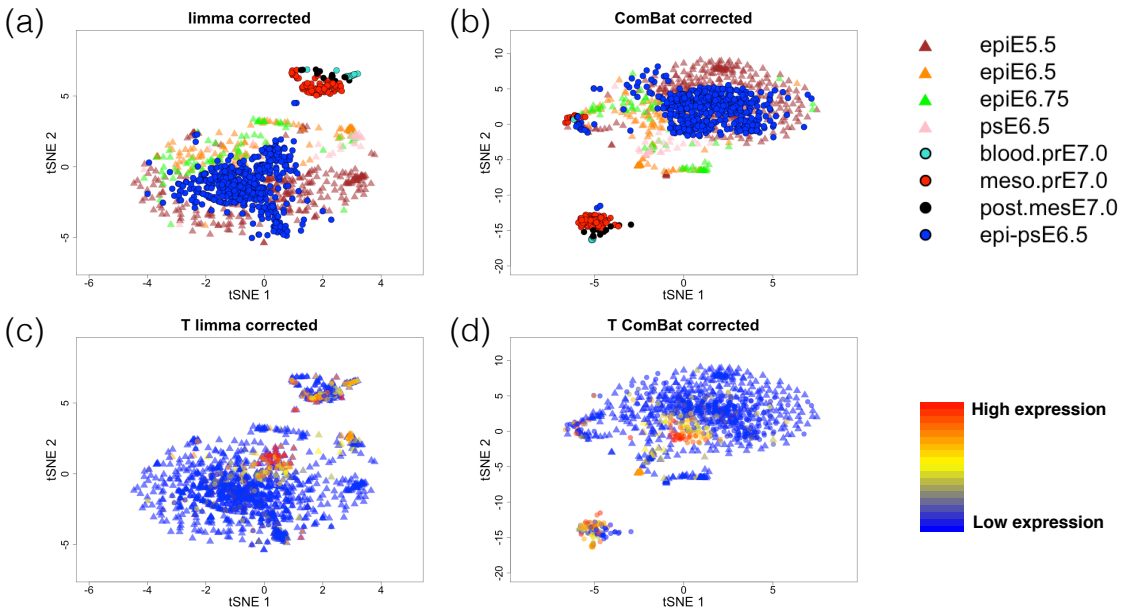
Supplementary Figure 2: PCA plots for shared cell types (GMP, CMP and MEP) between the two haematopoitic batches (shown by filled or open circles) for (a) uncorrected data and batch corrected with (b) MNN, (c) limma and (d) ComBat. Cells where the weights on PC2 was less than -0.1 are plotted at -0.1 for better visualization.



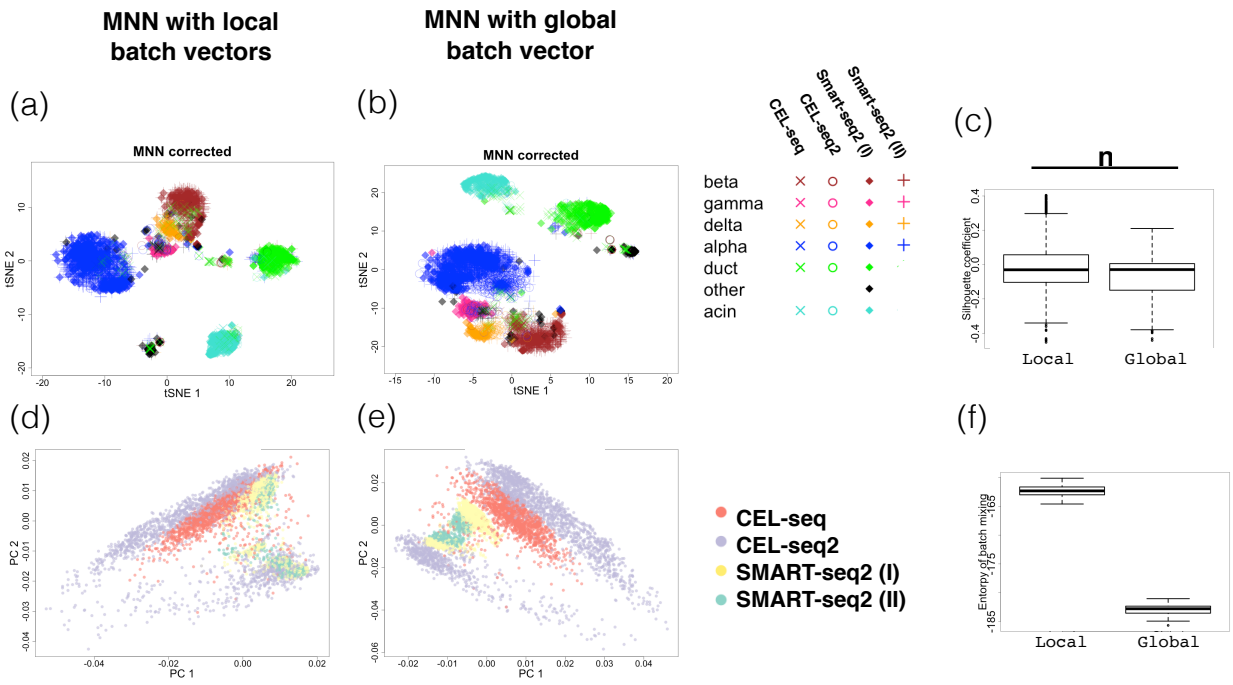
Supplementary Figure 3: PCA plots of pancreas data for (a) uncorrected (raw) and batch corrected data by (b) MNN, (c) limma and (d) ComBat coloured according to batch labels. (e) Boxplot of the total entropy of batch mixing on the first two PCs. Boxes indicate median and first and third quartile, whiskers extend to $\pm 1.5\times$ the interquartile ratio divided by the square root of the number of observations, and single points denote values outside this range. MNN and ComBat present significantly (p.value < 0.05 on Welch's t-test) larger total entropy of batch mixing compared to the uncorrected and limma corrected data.



Supplementary Figure 4: Re-clustering and differential expression analysis after MNN batch correction. Dimensionality reduction using t-SNE and partitioning into 8 clusters (k-means, k=8) on the MNN batch corrected gene expression values (a-top panel), followed by cell type label assignment using marker genes from GSE81076 (a-bottom panel). Combining data sets and correcting for batch effects increases statistical power to detect differentially expressed genes. Differential expression analysis between δ-islets (n=51 cells) and γ-islets (n=15 cells) within a single batch (GSE81076) finds 6 DE genes, including the two canonical marker genes *SST* and *PPY* (b-left panel). Increasing the number of cells by batch correction (δ-islets = 416, γ-islets = 407) allows the combination of datasets and improves the number of DE genes identified using an adjusted p-value threshold ($p<0.05$, purple dashed line; b-right panel). The DE genes identified with fewer cells are a subset of those identified with the larger sample size of combined data sets (c). The larger gene set identified marker genes (*SST, PPY*) as well as genes with functions involved in pancreatic islet development (*PAX6*), and δ-islet biology, e.g. *HADH* and *CD9*.



Supplementary Figure 5: t-SNE plots of scRNA-seq data of mouse cells during gastrulation, prepared in two laboratories (Mohammed *et al.* in press, triangles; Scialdone *et al.* 2016, circles). Plots were generated after (a) limma and (b) ComBat correction, with cells coloured by the cell type (epi: epiblast; ps: primitive streak; blood.pr: blood precursor, meso.pr: mesoderm precursor, post.mes: posterior mesoderm, epi-ps: epiblasts transitioning to the primitive streak). Cells were also coloured based on their expression of Brachyury (T), after (c) limma and (d) ComBat correction.



Supplementary Figure 6: t-SNE plots of pancreas data for batch corrected data by (a) MNN allowing for local batch vectors (default), (b) MNN with a single global batch vector for all cells. (c) Silhouette coefficients for clustering according to cell types for the two alternative settings of MNN. The difference between the Silhouette coefficients is nonsignificant (Welch's test pvalue=0.97) indicated by the symbol n over the boxplots. PCA plots of pancreas data for batch corrected data by (d) MNN allowing for local batch vectors (default), (e) MNN with a single batch vector for all cells. (f) Entropy of batch mixing on the first two PCs for batch corrected data with the two alternative settings of MNN. MNN with allowing for local batch vectors has significantly (Welch's test pvalue < 0.05) larger entropy compared to the global batch vector settings for MNN. Boxes indicate median and first and third quartile, whiskers extend to $\pm 1.5\times$ the interquartile ratio divided by the square root of the number of observations, and single points denote values outside this range.

Cell type / Platform	alpha	beta	gamma	delta	acid	duct	other/ unlabled	total
Cell-seq	196 (16%)	133 (11%)	15 (1%)	51 (4%)	257 (21%)	537 (45%)	0 (0%)	1189
Cell-seq2	961 (40%)	612 (25%)	98 (4%)	206 (9%)	280 (12%)	247 (10%)	0 (0%)	2404
Smart-seq2 (I)	886 (40%)	270 (12%)	197 (9%)	114 (5%)	185 (8%)	386 (18%)	150 (7%)	2188
Smart-seq2 (II)	838 (58%)	478 (33%)	82 (6%)	52 (4%)	0 (0%)	0 (0%)	0 (0%)	1450

Supplementary Table 1: The four pancreas data sets separated by the number of cells present in each cell type.