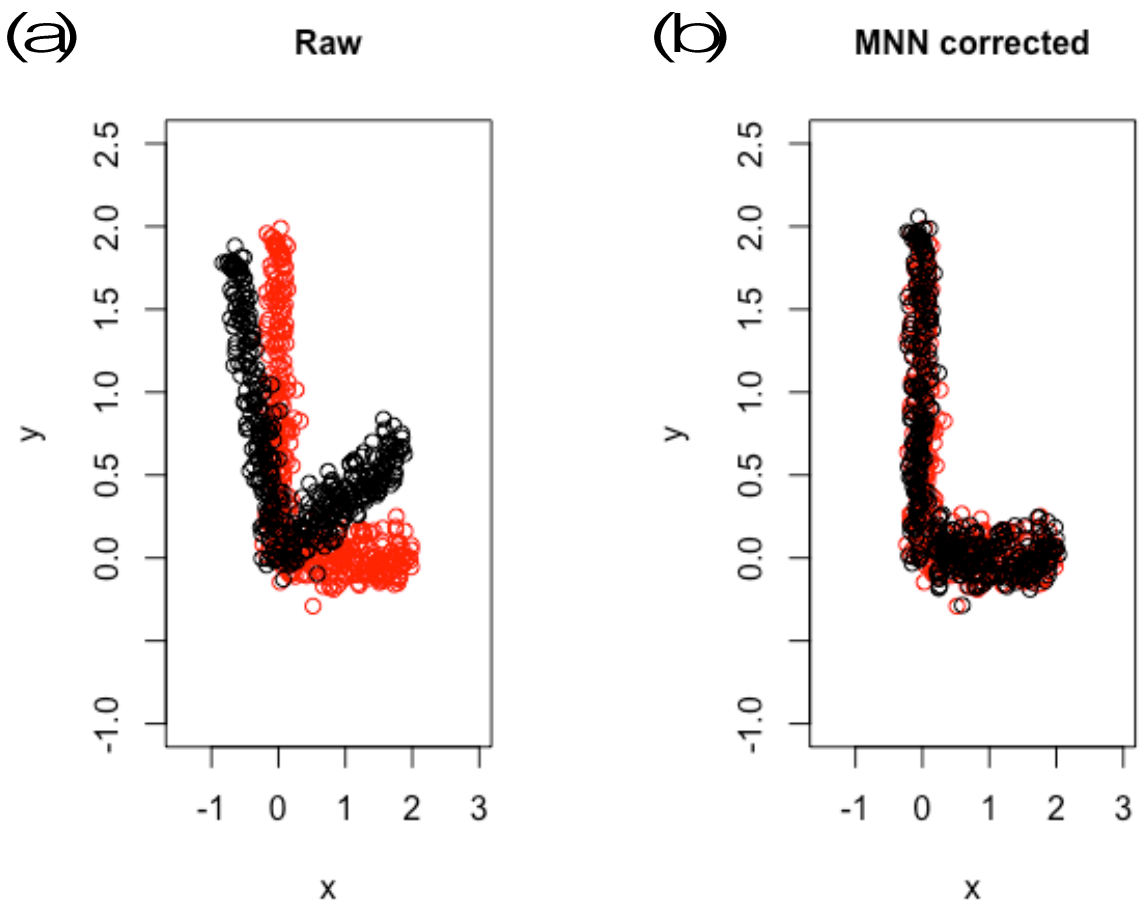
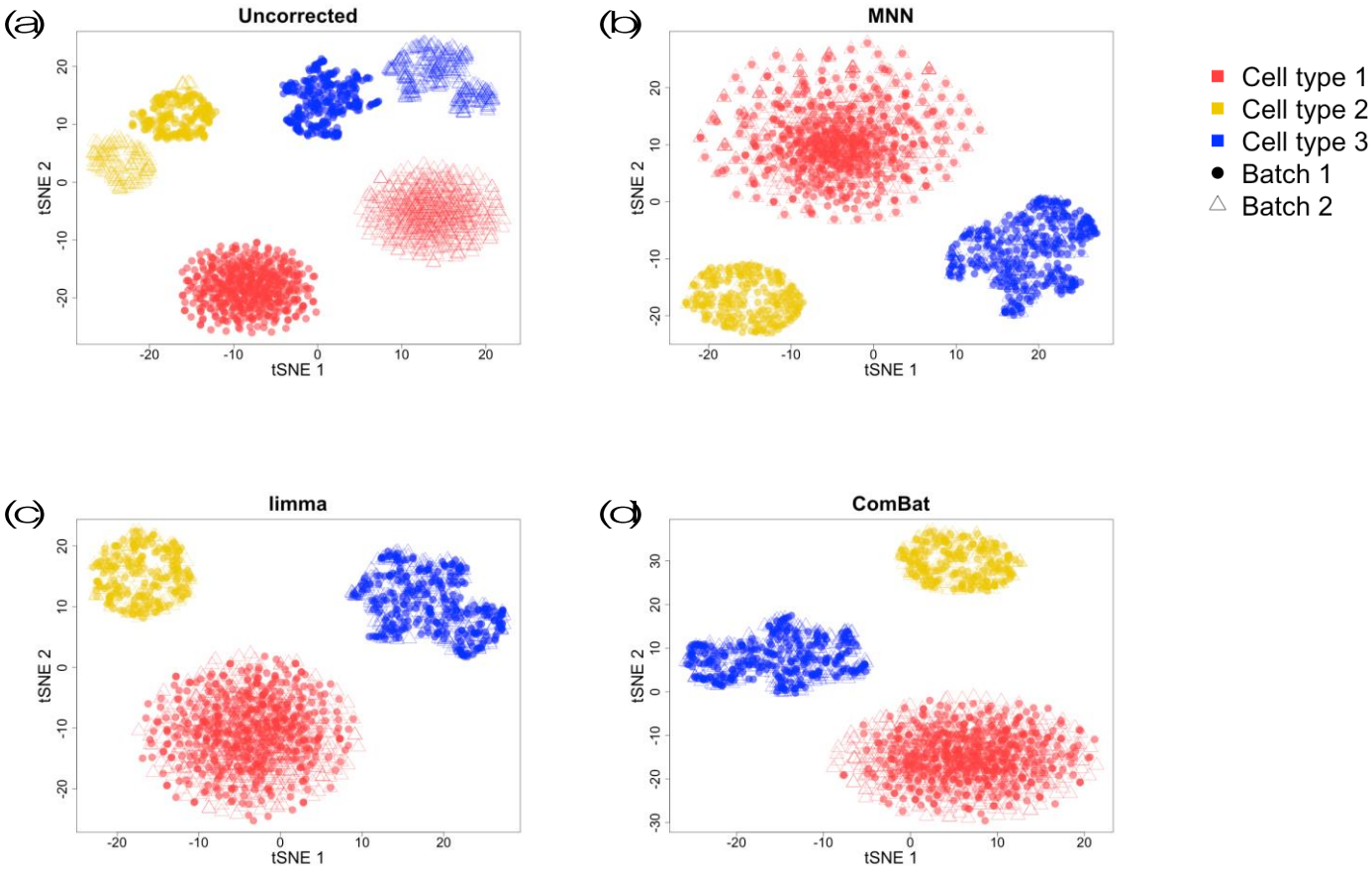


Supplementary Figure 1

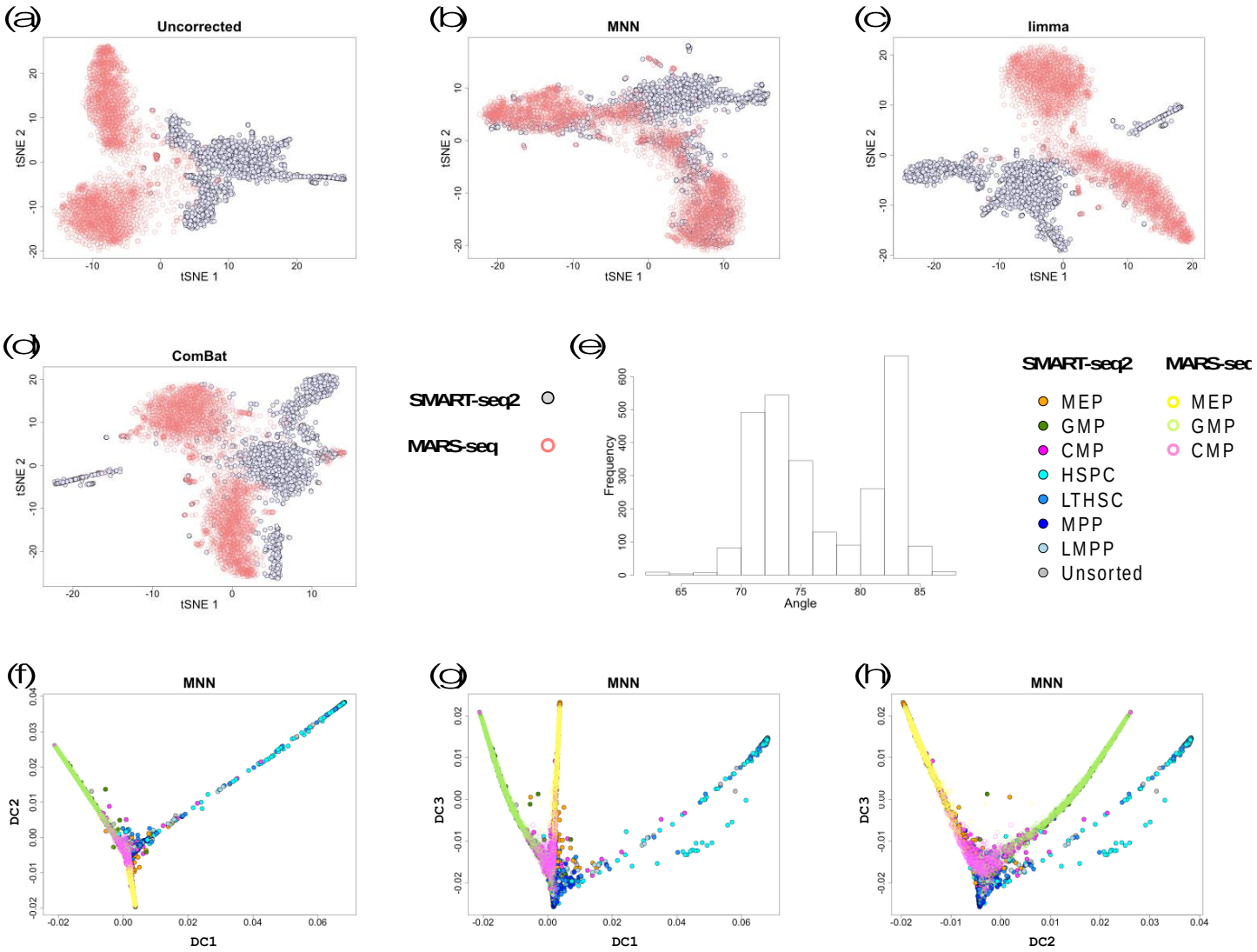
MNN corrects nonconstant batch effects.

By using locally linear corrections, MNN can handle non-constant batch effects, here simulated as a small angle rotation of data on two-dimensional x-y coordinates. Each shown batch contains 400 cells (points). The reference batch is shown in red and the second (rotated) batch is shown in black for (a) raw (uncorrected) data and (b) data after MNN correction.



Supplementary Figure 2

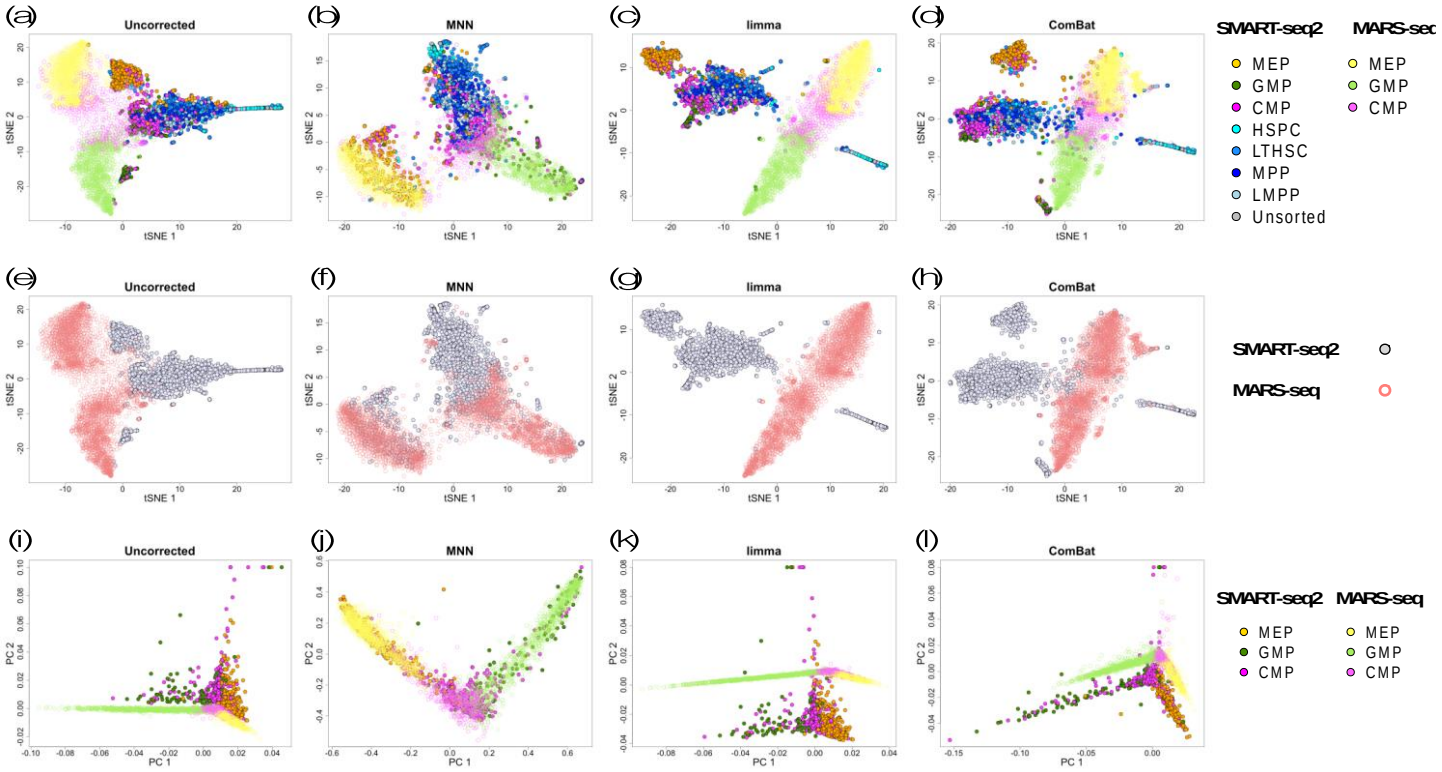
Simulation of batch effect in two batches with identical cell-type composition.
t-SNE plots of (a) the raw (uncorrected) simulated data, and the simulation data corrected by (b) our MNN approach, (c) limma and (d) ComBat. The filled circles and open triangles represent cells from the first and second batch respectively. The three different cell types are shown by different colours. 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 yield clusters consistent with the original simulated cell types. The data were simulated to have identical cell type compositions (0.2/0.3/0.5) in both batches, with each batch containing 1000 cells.



Supplementary Figure 3

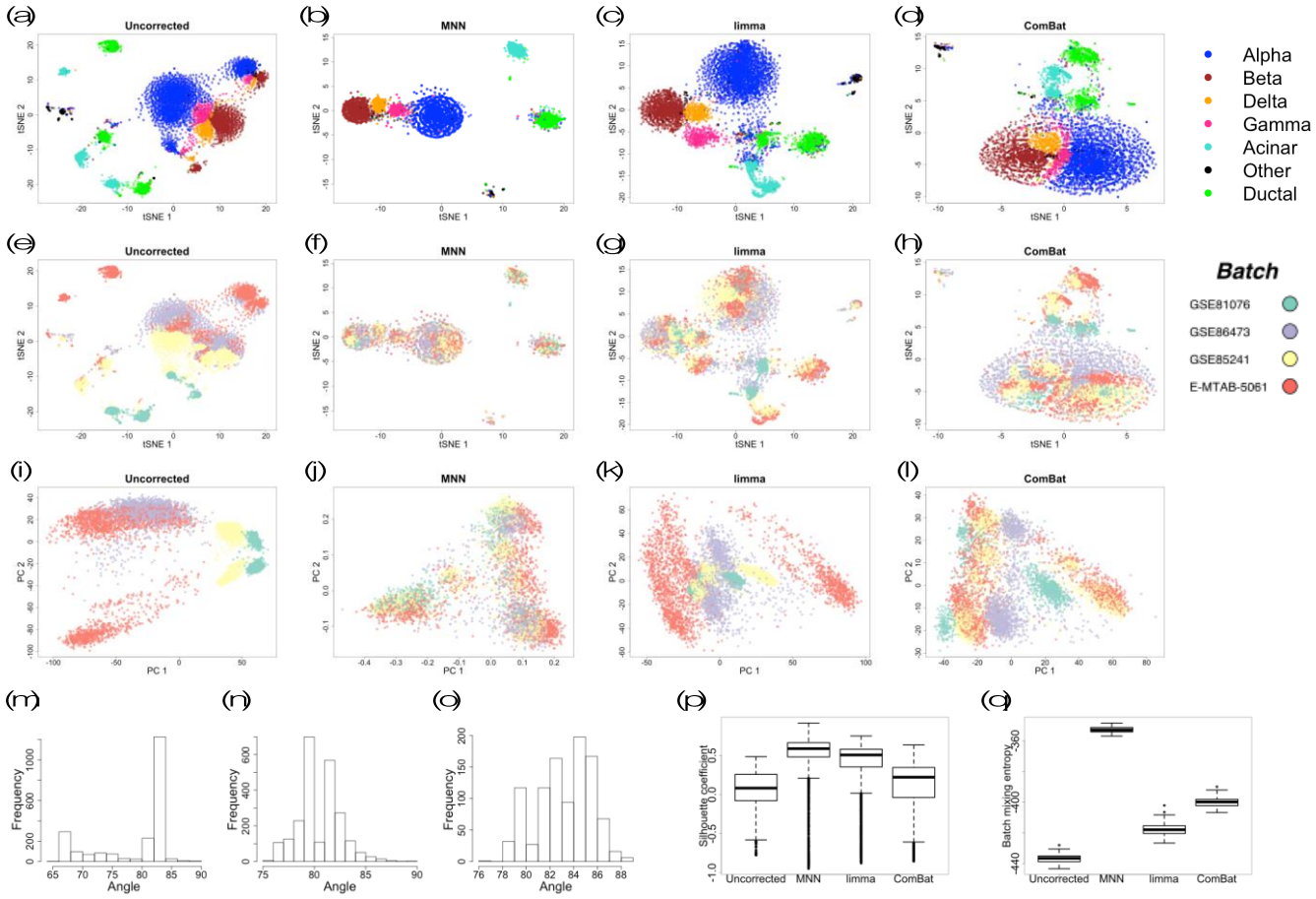
Analysis of the hematopoietic data by using all 3,904 highly variable genes.

t-SNE plots for (a) uncorrected data and data after correction by (b) our MNN approach, (c) limma and (d) ComBat. Cells are coloured according to their batch labels. (e) Histogram of the angle (f) between the first two SVD components of the reference data (SMART-seq2) and the correction vectors of the MARS-seq data calculated by MNN. Diffusion maps of the haematopoietic data after MNN correction, shown on the (f) first two diffusion components, (g) first and the third diffusion components, and (h) second and the third diffusion components.



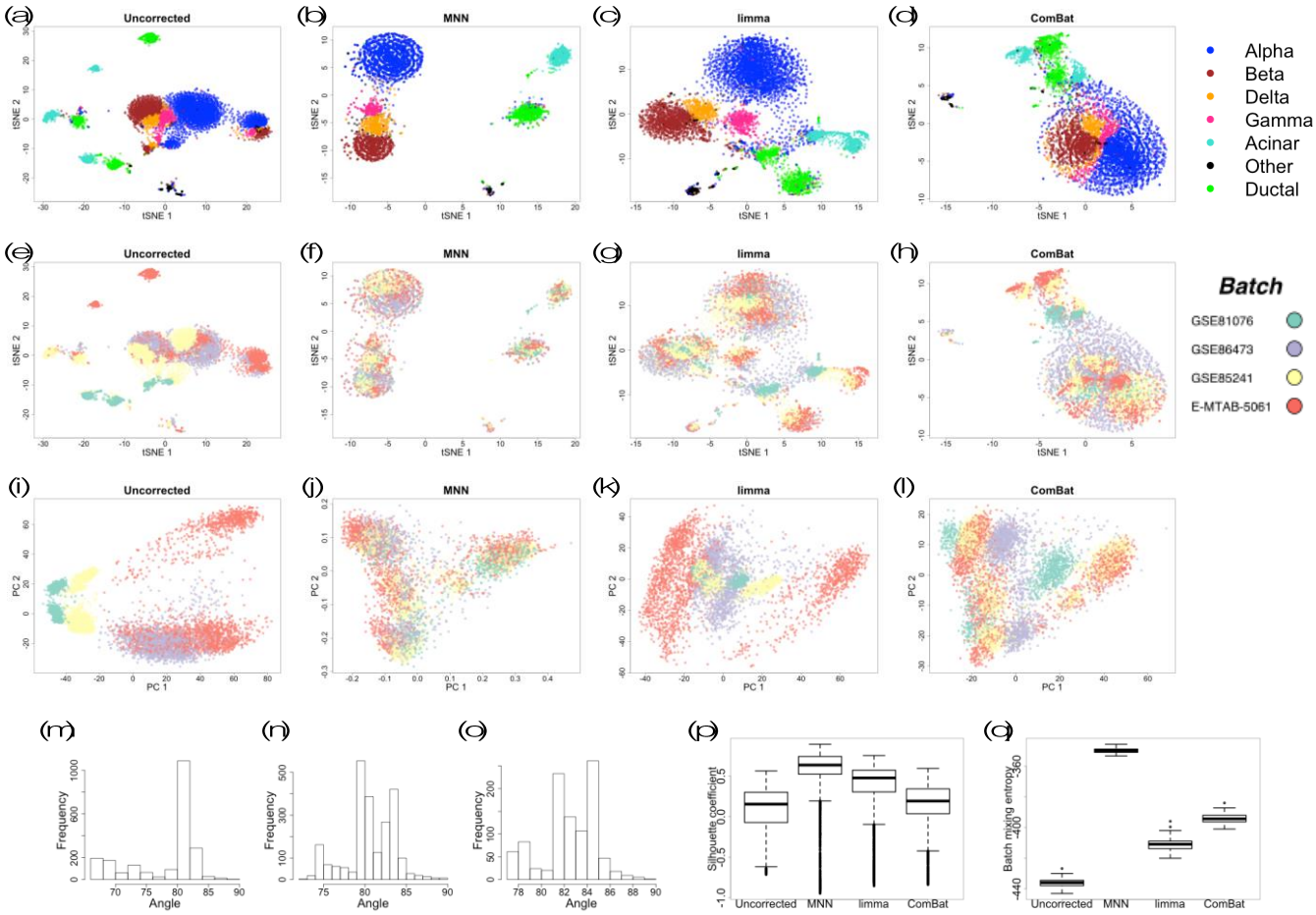
Supplementary Figure 4

Analysis of the hematopoietic data by using 1,500 genes randomly subsampled from the highly variable gene set.
t-SNE plots for (a) uncorrected data and data after correction by (b) MNN, (c) limma and (d) ComBat, coloured according to cell types. The same t -SNE plots are coloured according to batch for (e) uncorrected and batch-corrected data from (f) MNN, (g) limma and (h) ComBat. PCA plots for shared cell types (the SMART-seq2 batch with n=791 cells and the MARS-seq batch with n=2729 cells) between the two batches for (i) uncorrected data and batch-corrected data from (j) MNN, (k) limma and (l) ComBat.



Supplementary Figure 5

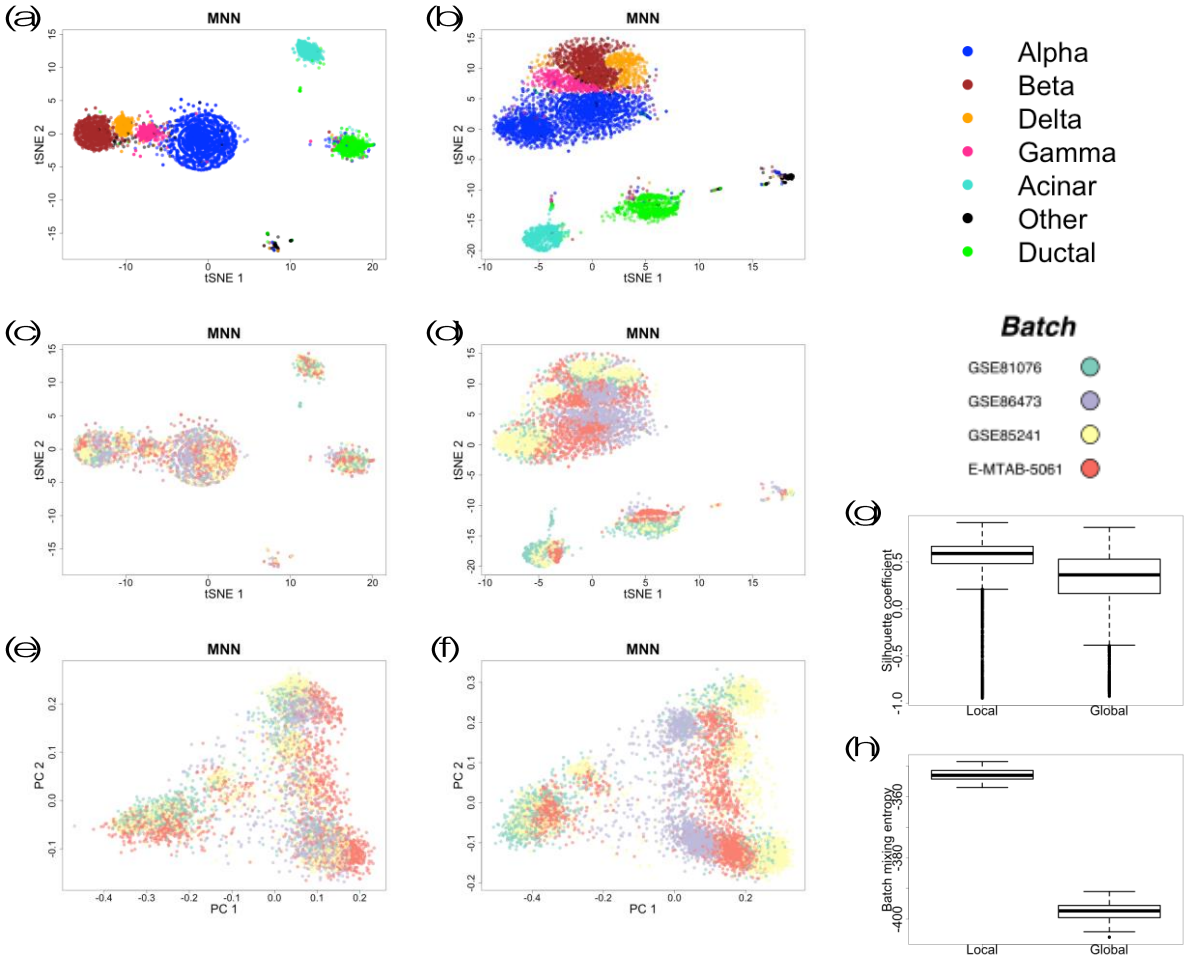
Analysis of the pancreas data by using all 2,507 highly variable genes.
t-SNE plots of (a) uncorrected data and data after correction by (b) MNN, (c) limma and (d) ComBat, coloured according to cell type labels. *t*-SNE plots were also generated for (e) uncorrected data and batch-corrected data from (f) MNN, (g) limma and (h) ComBat, coloured according to batch. PCA plots were also generated for (i) uncorrected and batch-corrected data from (j) MNN, (k) limma and (l) ComBat, coloured according to batch. Histograms of the angle between the batch effect vectors and the first two SVDs for the (m) reference (GSE85241) and the E-MTAB-5061 batch, (n) reference and the GSE86473 batch, and the (o) reference and the GSE81076 batch. (p) Silhouette coefficients according to cell type labels, with n=7096 (i.e. integrated number of cell from all four batches) observations for each boxplot. (q) Boxplots of the entropy of batch mixing on the first two PCs, with n=100 (i.e. number of bootstraps) observations for each boxplot. Boxes indicate median and first and third quartile, and whiskers extend to +/-1.5 times the interquartile ratio divided by the square root of the number of observations, and single points denote values outside this range.



Supplementary Figure 6

Analysis of pancreas data on 1,500 genes randomly subsampled from the highly variable gene set.

t-SNE plots of (a) uncorrected data and data corrected by (b) our MNN approach, (c) limma and (d) ComBat, coloured according to cell type labels. *t*-SNE plots of (e) uncorrected data and batch-corrected data from (f) MNN, (g) limma and (h) ComBat, coloured according to batch. PCA plots of (i) uncorrected data and batch-corrected data from (j) MNN, (k) limma and (l) ComBat, coloured according to batch. Histogram of the angle between the batch effect vectors and the first two SVD components for the (m) reference (GSE85241) and the E-MTAB-5061 batch, (n) reference and the GSE86473 batch, and the (o) reference and the GSE81076 batch. (p) Silhouette coefficients according to cell type labels, with n=7096 (i.e. integrated number of cell from all four batches) observations for each boxplot. (q) Boxplots of the entropy of batch mixing on the first two PCs, with n=100 (i.e. number of bootstraps) observations for each boxplot. Boxes indicate median and first and third quartile, and whiskers extend to +/-1.5 times the interquartile ratio divided by the square root of the number of observations, and single points denote values outside this range.



Supplementary Figure 7

Locally varying batch correction versus global (i.e., constant vector) batch correction.

t-SNE plots of pancreas data (GSE81076 with n=1007, GSE86473 with n= 2331, GSE85241 with n=1595 and E-MTAB-5061 with n=2163 cells) after batch correction with (a,c) MNN allowing for local batch vectors (default) or (b,d) MNN with a single global batch vector for all cells, coloured according to cell type labels (a,b) or batch (c,d). PCA plots of pancreas data after batch correction with (e) MNN allowing for local batch vectors (default) or (f) MNN with a single global batch vector for all cells, coloured according to batch. (g) Silhouette coefficients for clustering according to cell types after correction with two alternative settings of MNN, with n=7096 (i.e. integrated number of cell from all four batches) observations for each boxplot. The difference between the Silhouette coefficients is not significant (two-sided Welch's test p-value=0.97). (h) Entropy of batch mixing on the first two PCs for batch-corrected data with the two alternative settings of MNN, with n=100 (i.e. number of bootstraps) observations for each boxplot Allowing for local batch vectors has significantly (two-sided Welch's test p-value = 0.00001) larger entropy compared to the use of a global batch vector. Boxes indicate median and first and third quartile, and whiskers extend to +/-1.5 times the interquartile ratio divided by the square root of the number of observations, and single points denote values outside this range.

Supplementary Note 1: On the relation between MNN and projection methods

The existence of batch effects between data sets breaks the comparability of inter- to intra- data set cell similarities (e.g. correlation based measures, Euclidean distance), thus prohibiting generation of a combined reduced dimension map of the merged data. This problem can be overcome by first producing a reduced dimension map of a reference data set that includes all cell type present in the later data sets, before projecting the new data sets onto that reference map. Nestorowa et al. 2016 used data projection with diffusion maps (Angerer et al. 2015) to compare independently generated single-cell data sets and to develop new hypotheses for the study of haematopoietic stem cells. In another instance, Spitzer et al. 2015 used force-based dimension reduction and projection and showed that using a few landmark cell types from bone marrow (the most complete tissue in terms of covering several immune cell types) allowed mapping and comparison of compositions of immune cells across different tissues and different species. Such projection approaches require the reference batch to be a superset of all other batches, which precludes generation of an extensible map by integration of new data sets.

Our MNN approach relies on the same principle as projection methods, in that a given cell type in the reference batch is assumed to be more similar to cells of its own type rather than the other cell types present in the batch being compared to the reference. In Supplementary Note 4 we provide a mathematical proof for the conditions under which this property holds true. However, in contrast to projection methods, our MNN approach allows inclusion of new cell types in each batch, facilitating extension to previously unexplored spaces. Thus, extension of integrated maps by adding new populations is possible as long as there exists at least one shared population with the current map. While landmark projection approaches (based on finding nearest neighbours between data sets) have proved useful in the context of single-cell data, we present MNN as a novel type of non-linear projection that allows extension of the map through integration of new data, even when the reference batch is not a superset of later batches. We hypothesize that orthogonality of batch effects to the biological signal, together with slow variation of local batch vectors (also see Supplementary Note 4), explain the relative success of projection and landmark-based methods.

Supplementary Note 2: 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.

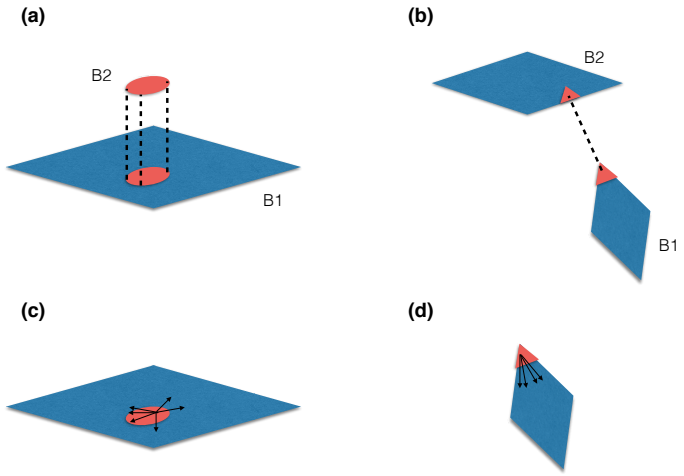
$$\begin{aligned} D^2(x,y) &= \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} - \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 = \left(\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||}\right)^2 + \left(\frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}\right)^2 - 2\frac{\mathbf{Y_x}}{||\mathbf{Y_x}||} \cdot \frac{\mathbf{Y_y}}{||\mathbf{Y_y}||}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

Supplementary Note 3: Violations of the MNN assumptions

Violation of any of the three assumptions of MNN may result in misinterpretation of data, as briefly discussed below.

(i) **There is at least one cell population that is present in both batches (i.e., in the reference and the new batch to be merged with it).** If this assumption does not hold, we cannot assume that there exists a single biological hyperplane (Matrix B in our math notation) that can biologically explain the data in both batches. In such cases, there is no point in merging the two batches. Nonetheless, if our approach is applied in this situation, we argue that MNN pairs

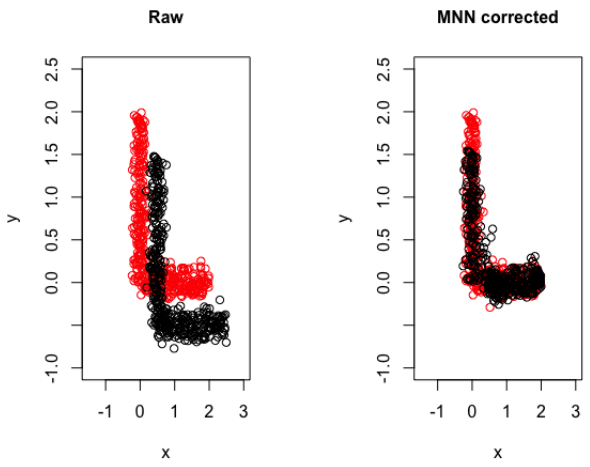
can only form on the "edges" of each batch. Subsequently, the MNN algorithm distinguishes edge cells by checking weather or not the vector sum of connecting vectors to a few nearest neighbours within the same batch has a non-zero phase (see the figure below). If all MNN pairs occur only on the edges, the algorithm recognizes this as a violation of this assumption and produces a warning message.



(a) Truly matching cell types between two related batches B1 and B2 are identified as MNN pairs; some MNN cells occur in the middle of batch B1. (b) For random floating objects (i.e. unrelated B1 and B2 batches) in high dimensions, spurious MNN pairs occur only on the "edges" of either of the batch manifolds. The vector sum of connecting vectors to a few nearest neighbours (within the same batch) has effectively a (c) zero phase for an MNN cell that is in the middle of the batch. (d) non-zero phase for an MNN cell that is near the edges of the batch.

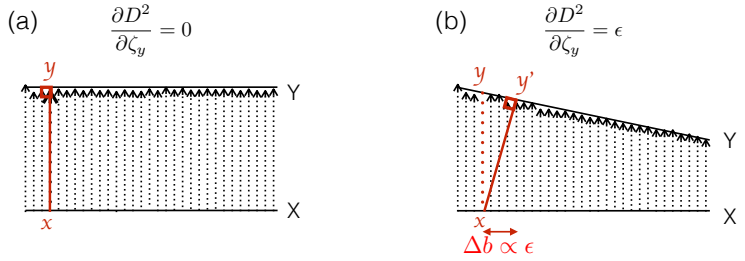
(ii) **The batch effect is almost orthogonal to the biological subspace.** While correlation between batch effects and biological signals may occur by chance in low dimensions, observation of such correlations are extremely unlikely in high dimensions. This explains why, following principal component analysis of RNA-Seq data, batch effects and biological signals generally appear as distinct orthogonal PCs (e.g. Figure 3f,h and i in the main text). The implementation of our MNN algorithm will report the angle between the calculated correction vectors and the first two SVD components of the reference batch, while merging each new batch. This facilitates checking of the validity of the orthogonality assumption (see Supplementary Figures 3i and 6m,n,o for example). If this assumption does not hold for a data set (e.g., when data are not high dimensional), matching cell types can not be identified accurately by MNN, which results in impaired performance as demonstrated in the figure below.

(iii) **Variation in the batch effects across cells is much smaller than the variation in the biological effects between different cell types.** The error in identification of biologically matching cells and thus the error of batch effect estimation is determined by how rapidly batch vectors vary on the biological hyperplane. Let $D(x, y)$ be the distance between two truly matching cells in batches X and Y, and ϵ represent how quickly $D(x, y)$ is changing along batch Y (i.e. the derivative with respect to the biological hyperplane parameter ζ_y). The figure below illustrates the dependence of the biological error range Δb on ϵ, meaning that if the batch effect variation rate along Y is too big, more biologically distant cells will be identified as MNN pairs. In other words, if biological variation is too small compared to batch effect variations (or noise), the biological distinctions between cells will be of a similar magnitude as the error in MNN pairs identification. In this case, application of MNN becomes redundant. This happens, for example, when one or more batches contain a high level of noise that dominates the biological signal. In addition, the



Biologically matching regions are not accurately identified as MNN pair cells, when the orthogonality assumption is violated. Here, the shift between the red and black batches is in the same plane as the L shape of each batch. As a result, the batch effect correction is inadequate.

assumption of slowly varying batch vectors allows for local estimation of batch vectors by Gaussian kernel averaging of MNN vectors (see Section 2 in the main text).



The error in identification of biologically similar cells via MNN's nearest neighbours approach (Δb) is proportional to the size of batch vector variations ϵ. Here X and Y present the batches and x and y are two truly matching cells in each batch. y' presents the incorrectly identified counterpart of x in the presence of batch effect variation.

A good check for validity of assumption iii is to run a SVD on each batch separately and make sure that the first SVD components indeed capture biological signals rather than noise. This assumption is particularly important for the batch that is used as the reference by MNN. For data sets with too few cells, for example, the biological hyperplane might not be well-defined in high dimensions and the biological signal might be dominated by noise. With its current implementation, MNN is unable to correct batch effect where the biological hyperplane is not distinguishable in the reference batch.

Supplementary Note 4: 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), they will be identified as mutual nearest neighbours of each other. A required assumption (assumption ii) is that of 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 signal B and a batch-specific signal W_b, 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 $x \in X, y \in Y$ is calculated as

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

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 weakly among neighbouring cell types, whereas variability due to the biological signal can be quite large (assumption iii). 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. Thus, cells with similar biological coefficients (e.g., same cell type/state) will be correctly identified as mutual nearest neighbours in our approach.

Nature Biotechnology: doi:10.1038/nbt.4091

Supplementary Note 5: The MNN batch correction algorithm

Algorithm 1

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 $N_1 + ... + N_n$ cells and G_I inquiry genes.

for each batch B_i **do**

- Cosine normalise the expression both data matrices of the highly variable genes and the inquiry genes.

end for

- Define the first batch (B_1) as the first reference (C). Consider $C_{HVG} \leftarrow B_{1,HVG}$ and $C_I \leftarrow B_{1,I}$ for the highly variable genes and inquiry genes space correspondingly.

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

- Using k nearest neighbouring cells search between the batches C and B_i, find corresponding MNN pairs l and m in the 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 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 $\vec{u}_{HVG}(x) = \frac{\sum_m \vec{v}_{HVG}^{ml} W_{HVG}(x,m)}{\sum_m W_{HVG}(x,m)}$.
- Correct for the batch effect in inquiry genes and 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 $B'_{i,HVG}$ to C_{HVG} (for the highly variable genes space) as well as $B'_{i,I}$ to C_I (for the inquiry genes space).

end for

The default setting of the parameters is as follows:

- The default number of nearest neighbours search for identification of MNN pairs is chosen $k = 20$ as good trade-off between sensitivity and tolerance to noise and (slight) non-orthogonality of batch effects, in the context of experiments involving 1000's of cells. We interpret $k = 20$ as the lower bound on the number of cells in each group of cells. $k = 1$ is too sensitive a choice; due to the random sampling of cells there is a very little chance that two cells find each other as the first rank nearest neighbour, which results in identification of too few MNNs. A larger k is more suitable, as it accounts for the noise in the data sets and allows for identification of cells with slightly differing biological coordinates between batches as MNNs. However, if k is chosen to be too large, the ability to accurately pair cells is compromised because cells from different populations may be considered as MNNs.

- We choose the default Gaussian kernel width $\sigma^2 = 1$ in the cosine normalized space, where the L^2 norm of any expression vector is equal to 1.

Nature Biotechnology: doi:10.1038/nbt.4091

Cell type→ Batch↓	Alpha	Beta	Gamma	Delta	Acinar	Ductal	Other/Unlabled	Total
GSE81076	178 (18%)	160 (16%)	24 (2%)	54 (5%)	238 (24%)	332 (33%)	21 (2%)	1007
GSE85241	852 (37%)	475 (20%)	119 (5%)	198 (85%)	288 (12%)	261 (11%)	138 (6%)	2331
GSE86473	944 (59%)	502 (31%)	92 (6%)	57 (4%)	0 (0%)	0 (0%)	0 (0%)	1595
E-MTAB-5061	874 (40%)	262 (12%)	194 (9%)	112 (5%)	185 (8%)	382 (18%)	154 (7%)	2163

Supplementary Table 1

The number of cells from each cell type present in the pancreas data sets.

Supplementary References

Angerer, Philipp et al. (2015). "destiny: diffusion maps for large-scale single-cell data in R". In: *Bioinformatics* 32.8, pp. 1241–1243.

Nestorowa, Sonia et al. (2016). "A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation". In: *Blood* 128.8, e20–e31.

Spitzer, Matthew H et al. (2015). "An interactive reference framework for modeling a dynamic immune system". In: *Science* 349.6244, p. 1259425.