Batch effects in single-cell RNA sequencing data are corrected by matching mutual nearest neighbours

Laleh Haghverdi[1,2], Aaron T. L. Lun[3], Michael D. Morgan[4], John C. Marioni[1,3,4]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom.
[2] Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany.
[3] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom.
[4] Wellcome Trust Sanger Institute, Cambridge, United Kingdom.

Correspondence should be addressed to J.C.M.: John.Marioni@cruk.cam.ac.uk

Editor Summary: Differences in gene expression between individual cells of the same type are measured across batches and used to correct technical artefacts in single-cell RNA sequencing data

Large-scale single-cell RNA sequencing (scRNA-seq) datasets that are produced in different laboratories and at different times contain batch effects that could compromise integration and interpretation of these data. Existing scRNA-seq analysis methods incorrectly assume that the composition of cell populations is either known, or the same, across batches. We present a strategy for batch correction that is based on the detection of mutual nearest neighbours (MNN) in the high-dimensional expression space. Our approach does not rely on pre-defined or equal population compositions across batches, and only requires that a subset of the population be shared between batches. We demonstrate the superiority of our approach over existing methods using both simulated and real scRNA-seq data sets. Using multiple droplet-based scRNA-seq data sets, we demonstrate that our MNN batch-effect correction method scales to large numbers of cells.

INTRODUCTION

The decreasing cost of single-cell RNA sequencing experiments [1] [2] [3] [4] has encouraged the establishment of large-scale projects such as the Human Cell Atlas, which profile the transcriptomes of thousands to millions of cells. For such large studies, logistical constraints inevitably dictate that data are generated separately i.e., at different times and with different operators. Data may also be generated in multiple laboratories using different cell dissociation and handling protocols, library preparation technologies and/or sequencing platforms. All of these factors result in batch effects [5] [6] , where the expression of genes in one batch differs systematically from those in another batch. Such differences can mask underlying

biology or introduce spurious structure in the data, and must be corrected prior to further analysis to avoid misleading conclusions.

Most existing methods for batch correction are based on linear regression. The limma package provides the *removeBatchEffect* function [7], which fits a linear model containing a blocking term for the batch structure to the expression values for each gene. Subsequently, the coefficient for each blocking term is set to zero and the expression values are computed from the remaining terms and residuals, yielding a new expression matrix without batch effects. The ComBat method [8] uses a similar strategy but performs an additional step involving empirical Bayes shrinkage of the blocking coefficient estimates. This stabilizes the estimates in the presence of limited replicates by sharing information across genes. Other methods such as RUVseq [9] and svaseq [10] are also frequently used for batch correction, but focus primarily on identifying unknown factors of variation, e.g., due to unrecorded experimental differences in cell processing. Once these factors are identified, their effects can be regressed out as described previously.

Existing batch correction methods were specifically designed for bulk RNA-seq. Thus, their applications to scRNA-seq data assume that the composition of the cell population within each batch is identical. Any systematic differences in the mean gene expression between batches are attributed to technical differences that can be regressed out. However, in practice, population composition is usually not identical across batches in scRNA-seq studies. Even assuming that the same cell types are present in each batch, the abundance of each cell type in the data set can change depending upon subtle differences in cell culture or tissue extraction, dissociation and sorting, etc. Consequently, the estimated coefficients for the batch blocking factors are not purely technical, but contain a non-zero biological component due to differences in composition. Batch correction based on these coefficients will thus yield inaccurate representations of the cellular
expression proles, potentially yielding worse results than if no correction was performed.

An alternative approach for data merging and comparison in the presence of batch effects uses a set of landmarks from a reference data set to project new data onto the reference [11] [12]. The rationale here is that a given cell type in the reference batch is most similar to cells of its own type in the new batch. Such projection strategies can be applied using several dimensionality reduction methods such as principal components analysis (PCA), diffusion maps or by force-based methods such as t-distributed stochastic nearest-neighbour embedding (*t*-SNE). This strategy depends on the selection of landmark points in high dimensional space picked from the reference data set, which cover all cell types that might appear in the later batches. However, if the new batches include cell types that fall outside the transcriptional space explored in the reference batch, these cell types will not be projected to an appropriate position in the space defined by the landmarks (Supplementary Note 1).

Here, we propose a new method for removal of discrepancies between biologically related batches based on the presence of mutual nearest neighbours (MNNs) between batches, which are considered to define the most similar cells of the same type across batches. The difference in expression values between cells in a MNN pair provides an estimate of the batch effect, which is made more precise by averaging across many such pairs. A correction vector is obtained from the estimated batch effect and applied to the expression values to perform batch correction. Our approach automatically identifies overlaps in population composition between batches and uses only the overlapping subsets for correction, thus avoiding the assumption of equal composition required by other methods. We demonstrate that our approach outperforms existing methods on a range of simulated and real scRNA-seq data sets involving different biological systems and technologies.

RESULTS

Matching mutual nearest neighbours for batch correction

Our approach identifies cells between different experimental batches or replicates that have mutually similar expression profiles. We infer that any differences between

these cells in the high-dimensional gene expression space are driven by batch effects (i.e., technical differences induced by the operator or other experimental artefacts) and do not represent the underlying biology of interest. We note that our definition of a batch effect may also incorporate some signal driven by biological features that are not of interest (e.g., differences between samples due to genotype). Upon correction, multiple batches can be “joined up" into a single data set (Figure 1a).

The first step of our method involves global scaling of the data using a cosine normalization. More precisely, if $\mathrm{Y_x}$ is the expression vector for cell x, we define the cosine normalization as:

$$Y_x \leftarrow \frac{Y_x}{\| Y_x \|} \qquad (1)$$

Subsequently, we compute the Euclidean distance between the cosine-normalized expression profiles of pairs of cells. Calculating Euclidean distances on this normalised data is equivalent to the use of cosine distances on the original expression values (Supplementary Note 2). Cosine distances have been widely used for measuring cell similarities based on their expression profiles [11] [13] [14] [15] and are appealing as they are scale-independent [15], which makes them robust to technical differences in sequencing depth and capture efficiency between batches.

The next step involves identification of **mutual nearest neighbours**. Consider a scRNA-seq experiment consisting of two batches 1 and 2. For each cell $\mathrm{i_1}$ in batch 1, we find the k cells in batch 2 with the smallest distances to $\mathrm{i_1}$, i.e., its k nearest neighbours in batch 2. We do the same for each cell in batch 2 to find its k nearest neighbours in batch 1. If a pair of cells from each batch are contained in each other's set of nearest neighbours, those cells are considered to be mutual nearest neighbours (Figure 1b). We interpret these pairs as containing cells that belong to the same cell type or state, despite being generated in different batches. This means that any systematic differences in expression level between cells in MNN pairs should represent the batch effect.

Our use of MNN pairs involves three assumptions: (i) there is at least one cell population that is present in both batches, (ii) the batch effect is almost orthogonal to the biological subspace, and (iii) batch effect variation is much smaller than the biological effect variation between different cell types (see Supplementary Note 3 for a more detailed discussion of these assumptions). The biological subspace refers to a set of basis vectors, each of length equal to the number of genes, which represent biological processes. For example, some of these vectors may represent the cell cycle; some vectors may define expression profiles specific to each cell type; while other vectors may represent differentiation or activation states. The true expression profile of each cell can be expressed as a linear sum of these vectors. Meanwhile, the batch effect is represented by a vector of length equal to the number of genes, which is added to the expression profile for each cell in the same batch. Under our assumptions, it is straightforward to show that cells from the same population in different batches will form MNN pairs (Supplementary Note 4). This can be more intuitively understood by realizing that cells from the same population in different batches form parallel hyperplanes with respect to each other (Figure 1b). We also note that the orthogonality assumption is weak for a random one-dimensional batch effect vector in high-dimensional data, especially given that local biological subspaces usually have much lower intrinsic dimensionality than the total number of genes in the data set.

For each MNN pair, a pair-specific batch correction vector is computed as the vector difference between the expression profiles of the paired cells. While a set of biologically relevant genes (e.g. highly variable genes) can facilitate identification of MNNs, the calculation of batch vectors does not need to be performed in the same space. Therefore, we can calculate the batch vectors for a different set of inquiry genes (Supplementary Note 5). A cell-specific batch correction vector is then calculated as a weighted average of these pair-specific vectors, computed using a Gaussian kernel. This approach stabilizes the correction for each cell and ensures that it changes smoothly between adjacent cells in the high-dimensional

expression space. This Gaussian smoothing of batch vectors enables a locally linear batch correction, i.e., each MNN pair batch vector will contribute to the batch effect for cells in the neighbourhood of the corresponding pair within each batch. Such locally linear correction of batch effects results in an overall correction that can tolerate non-constant batch effects (Supplementary Figure 1). We emphasize that this correction is performed for all cells, regardless of whether or not they participate in a MNN pair. This means that correction can be performed on all cells in each batch, even if they do not have a corresponding cell type in the other batches.

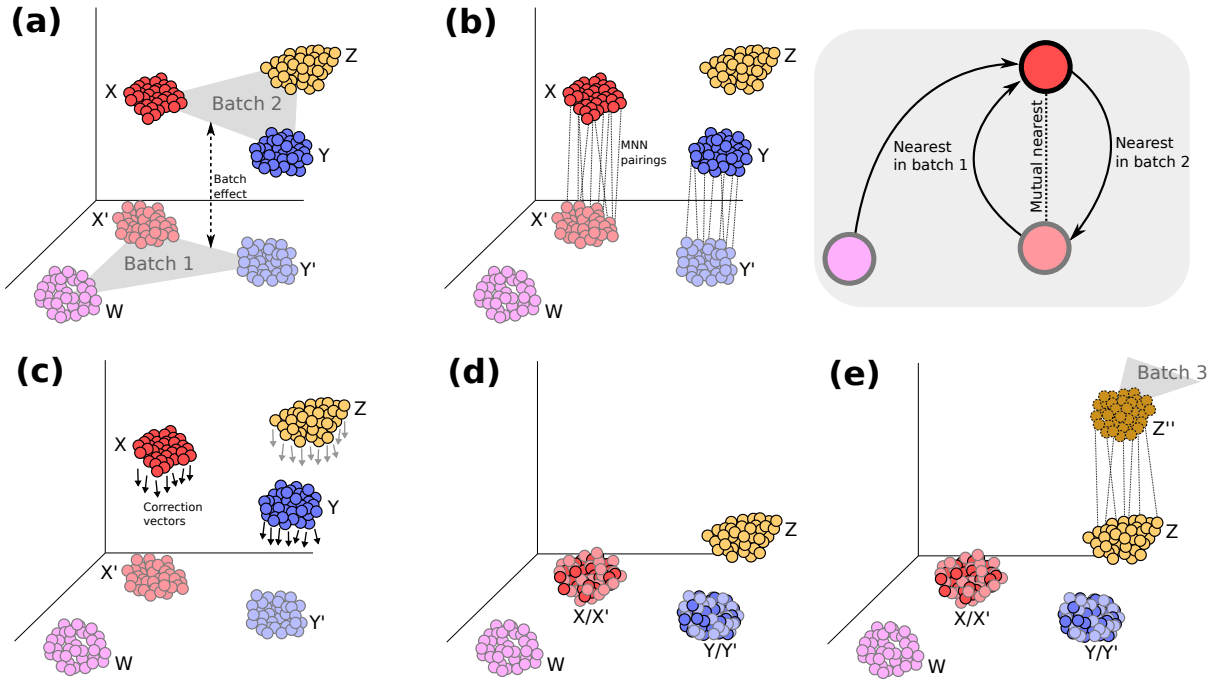

Figure 1: Schematics of batch effect correction by MNN. (a) Batch 1 and batch 2 in high dimensions with an almost orthogonal batch effect difference between them. (b) The algorithm identifies matching cell types by finding mutual nearest neighbouring pairs of cells (grey box). (c) Batch correction vectors are calculated between the MNN pairs. (d) Batch 1 is regarded as the reference and batch 2 is integrated into it by subtraction of correction vectors. (e) The integrated data are considered as the reference and the procedure is repeated for integration of any new batch.

MNN correction outperforms existing methods on simulated data

We generated simulated data for a simple scenario with two batches of cells, each consisting of varying proportions of three cell types (Online Methods). We applied

each batch correction method – our MNN-based correction method, limma and ComBat – to the simulated data, and evaluated the results by inspection of *t*-SNE plots [16] (Online Methods). Proper removal of the batch effect should result in the formation of three clusters, one for each cell type, where each cluster contains a mixture of cells from both batches. However, we only observed this ideal result after MNN correction (Figure 2). Expression data that were uncorrected or corrected with the other methods exhibited at least one cluster containing cells from only a single batch, indicating that the batch effect was not fully removed. This is fully attributable to the differences in population composition, as discussed earlier. Repeating the simulation with identical proportions of all cell types in each batch yielded equivalent performance for all methods (Supplementary Figure 2).

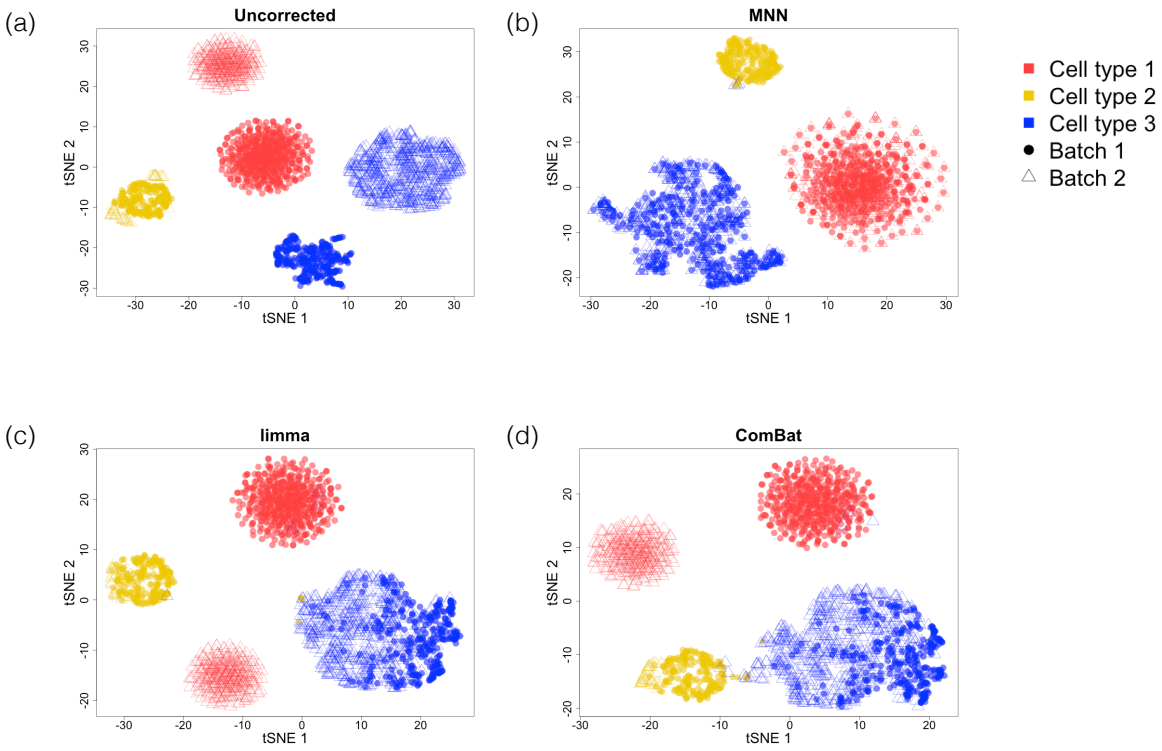

Figure 2: *t*-SNE plots of simulated scRNA-seq data containing two batches of different cell types (with each batch containing n=1000 cells), (a) before and after correction with (b) our MNN method, (c) limma or (d) ComBat. In this simulation, each batch (closed circle or open triangle) contained different numbers of cells in each of three cell types (specified by colour).

MNN correction outperforms existing methods on haematopoietic data

To demonstrate the applicability of our method on real data, we considered two haematopietic data sets generated in different laboratories using two different scRNA-seq protocols. In the first data set [12], the authors used the SMART-seq2 protocol [17] to profile single cells from haematopoietic stem and progenitor cell (HSPC) populations in 12-week-old female mice. Using marker expression profiles from fluorescence-activated cell sorting (FACS), known cell type labels were retrospectively assigned to cells (Online Methods). This included multipotent progenitors (MPP), lymphoid-primed multipotent progenitors (LMPP), haematopoietic stem and progenitor cells (HSP), haematopoietic stem cells (HSC), common myeloid progenitors (CMP), granulocyte-monocyte progenitors (GMP), and megakaryocyte-erythrocyte progenitors (MEP). In the second data set [18], the authors used the MARS-seq protocol to assess single-cell heterogeneity in myeloid progenitors for 6- to 8-week-old female mice. Again, indexed FACS was used to assign a cell type label (MEP, GMP or CMP) to each cell.

To assess performance, we performed *t*-SNE dimensionality reduction on the expression data of the highly variable genes, before and after correction using each of the three methods (MNN, limma and ComBat) (Figure 3, a-d and Online Methods). Only MNN correction was able to correctly merge the cell types that were shared between batches, i.e., CMPs, MEPs and GMPs, while preserving the underlying differentiation hierarchy [12] [18] (Figure 3e). In contrast, the shared cell types still clustered by batch after correction with limma or ComBat, indicating that the batch effect had not been completely removed (see Supplementary Figure 3 for colouring by batch). This is attributable to the differences in cell type composition between batches, consistent with the simulation results. To ensure that these results were not due to an idiosyncrasy of the *t*-SNE method, we repeated our analysis with an alternative dimensionality reduction approach (PCA) using only the common cell types between the two batches (Figure 3 f-i). MNN correction was still the most effective at removing the batch effect compared to the other methods.

As a justification for the orthogonality of batch effect to the biological hyperplane, we present a histogram of the angle between the batch vectors calculated by MNN and

the first two singular value decomposition (SVD) components of the reference batch used in MNN (i.e., the SMART-seq2 data set). Most angles are close to 90°, supporting the near-orthogonality assumption (Supplementary Figure 3 e). A diffusion map [19] of the MNN corrected data (Supplementary Figure 3 f-h) shows the same differentiation hierarchy of cell types as observed in Figure 3e. Repeating the same analysis on a subset of randomly sampled genes (1500 out of the total of 3904 highly variable genes), yielded similar results, thus demonstrating the robustness of our analysis with respect to the input gene set (Supplementary Figure 4).

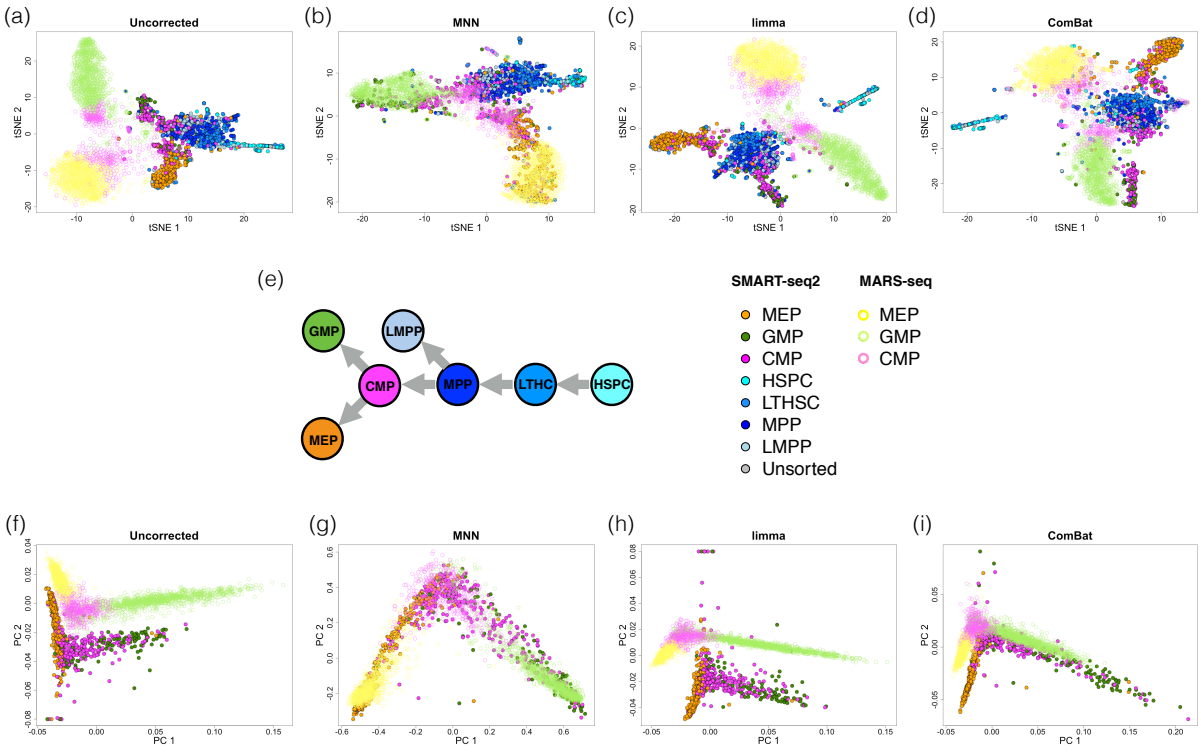

Figure 3: *t*-SNE plots of scRNA-seq count data for cells from the haematopoietic lineage, prepared in two batches using different technologies (SMART-seq2 with n=1920 cells, closed circle; MARS-seq, with n=2729 cells, open circle). Plots were generated (a) before and after batch correction using (b) our MNN method, (c) limma or (d) ComBat. Cells are coloured according to their annotated cell type. (e) The expected hierarchy of haematopoietic cell types. PCA plots of scRNA-seq count data for common cells types between the two batches of the haematopoietic lineage generated (SMART-seq2 with n=791 cells and MARS-seq, with n=2729 cells) (f) before and after batch correction using (g) our MNN method, (h) limma or (i) ComBat.

MNN correction outperforms existing methods on a pancreas data set

We further tested the ability of our method to combine more complex data sets generated using a variety of different methods. Here, we focused on the pancreas as it is a highly heterogeneous tissue with several well-defined cell types. We combined scRNA-seq data on human pancreas cells from four different publicly available data sets [20] [21] [22] [23], generated with two different scRNA-seq protocols (SMART-seq2 and CELseq/ CEL-seq2). Cell type labels were taken from the provided metadata, or derived by following the methodology described in the original publication (see Online Methods for further details of data preprocessing).

We applied MNN, limma and ComBat to the combined data set and examined the corrected data. All three batch correction methods improve the grouping of cells by their cell type labels (Online Methods, Supplementary Figure 5a-d). This is not surprising, as the discrepancy between cell type composition in the four batches is modest (Supplementary Table 1). However, even a small difference in composition is sufficient to cause ductal and acinar cells to be incorrectly separated following correction with limma or ComBat. By comparison, both cell types are coherently grouped across batches following MNN correction, consistent with the simulation results. To determine the effect of correction on the quality of cell type-based clustering, we assessed cluster separation by computing the average Silhouette widths for each cell type (Supplementary Figure 5, Online Methods). The average Silhouette coefficient after MNN correction is significantly larger than those in the uncorrected, limma and ComBat-corrected data ($p < 0{:}05$, two-sided Welch's *t*-test). Thus, MNN correction is able to reduce the between-batch variance within each cell type while preserving differences between cell types. We also computed the entropy of mixing (Online Methods) to quantify the extent of intermingling of cells from different batches. Batch corrected data using MNN show higher entropy of mixing compared to the uncorrected data and corrected data using limma or ComBat (Supplementary Figure 5). The improvement in the mixing of batches is observed in the reduced dimension space by either *t*-SNE or PCA (Supplementary Figure 5e-l). We again illustrate our assumption that batch effects are adequately removed when they lie orthogonally to the biological subspace (Supplementary Figure 5m-o). The observed structure in the pancreas data is robust to the size of the input gene set,

demonstrated by random subsampling of the total highly variable gene set (Supplementary Figure 6).

MNN correction improves differential expression analyses

Once batch correction is performed, the corrected expression values can be used in routine downstream analyses such as clustering and differential gene expression identification. To demonstrate, we used the MNN-corrected expression matrix to simultaneously cluster cells from all four pancreas data sets. Our new cluster labels were in agreement with the previous cell type assignments based on the individual batches, with an adjusted Rand index of 0.94 (a Rand index of 0 is equivalent to a random assignment, whilst a Rand index of 1 denotes a perfect match between previous and new assignments). Importantly, we obtained clusters for all batches in a single clustering step. This ensures that the cluster labels are directly comparable between cells in different batches. In contrast, if clustering were performed separately in each batch, there is no guarantee that a (weakly-separated) cluster detected in one batch has a direct counterpart in another batch.

We used our new clusters to perform a differential expression (DE) analysis between the δ-islet cluster and the γ-islet cluster. Using cells from all batches, we detected 76 differentially expressed genes at a false discovery rate (FDR) of 5% (Figure 4c). This set included the marker genes for the cells included in the analysis (*PPY*, *SST*), genes involved in pancreatic islet cell development (*PAX6*) and genes recently implicated in δ-islet function and type 2 diabetes development (*CD9*, *HADH*) [22]. For comparison, we repeated the DE analysis using only cells from each batch in which both cell types were present [21] [22] [20]. This yielded only 12, 59 and 88 genes respectively, at a FDR of 5%, which encompass 14.5-57.9% of those detected using all cells (Figure 4d). Merging data sets is beneficial as it increases the number of cells without extra experimental work; improves statistical power for downstream analyses such as differential gene expression; and in doing so, provides additional biological insights. To this end, our MNN approach is critical as it ensures that merging is performed in a coherent manner.

MNN correction is applicable to droplet RNA-seq technology

The advent of droplet-based cell capture, lysis, RNA reverse transcription and subsequent expression profiling by sequencing has allowed single cell expression experiments to be scaled up to tens and hundreds of thousands of cells [2] [3] [24]. These technologies are ideal for testing the scalability and applicability of our correction method to large scRNA-seq data sets. We specifically applied our MNN approach to two large data sets of droplet-based scRNA-seq derived from the commercial 10X Genomics Chromium platform [24]. We selected data sets in which there were a mixture of cell identities and complexities; namely 68,000 peripheral blood mononuclear cells (PBMCs) and 4,000 T cells, derived from different donors. PBMCs contain a milieu of peripheral adaptive and innate immune white blood cells as they circulate through the human vasculature, while peripheral T cells contain a mixture of naïve and antigen-exposed lymphocytes involved in active immune surveillance.

A naive merging of these two data sets without accounting for batch effects illustrates the separation of the T cells from their counterparts in the PBMC data (Figure 5a,b). Combination of these two data sets using MNNs demonstrates that the separate peripheral T cells map to the T cell subsets within the PBMC mixture (Figure 5c,d). Importantly, other peripheral lymphocyte relationships are not distorted by the correction applied, despite the absence of MNNs in the T cell data set (Figure 5c). Specifically, we note that 4446/4459 (99.7%) of individual T cells map onto their appropriate counterparts in the PBMC data set (Figure 5). The remaining 13/4459 (0.3%) map primarily to a small cluster of unknown ontogeny and to the edges of a large cluster of monocytes. Conversely, 14 non-T cells (0.3%; specifically monocytes) mapped to T cell clusters inappropriately.

As the size of single cell expression data sets increases, there will be a growing need for computational methods that can scale up to meet these requirements. To demonstrate the scalability of our method, we sampled different proportions of cells

from the 68K PBMC data set, and corrected the batch effect between each subsample and the 4K T cell data. Within the range of 7,000 to 70,000 cells we see an approximately linear time increase (Figure 5e). This demonstrates that our method is applicable to both the nature of droplet technology-derived single cell expression data, and the scale of current and future data sets.

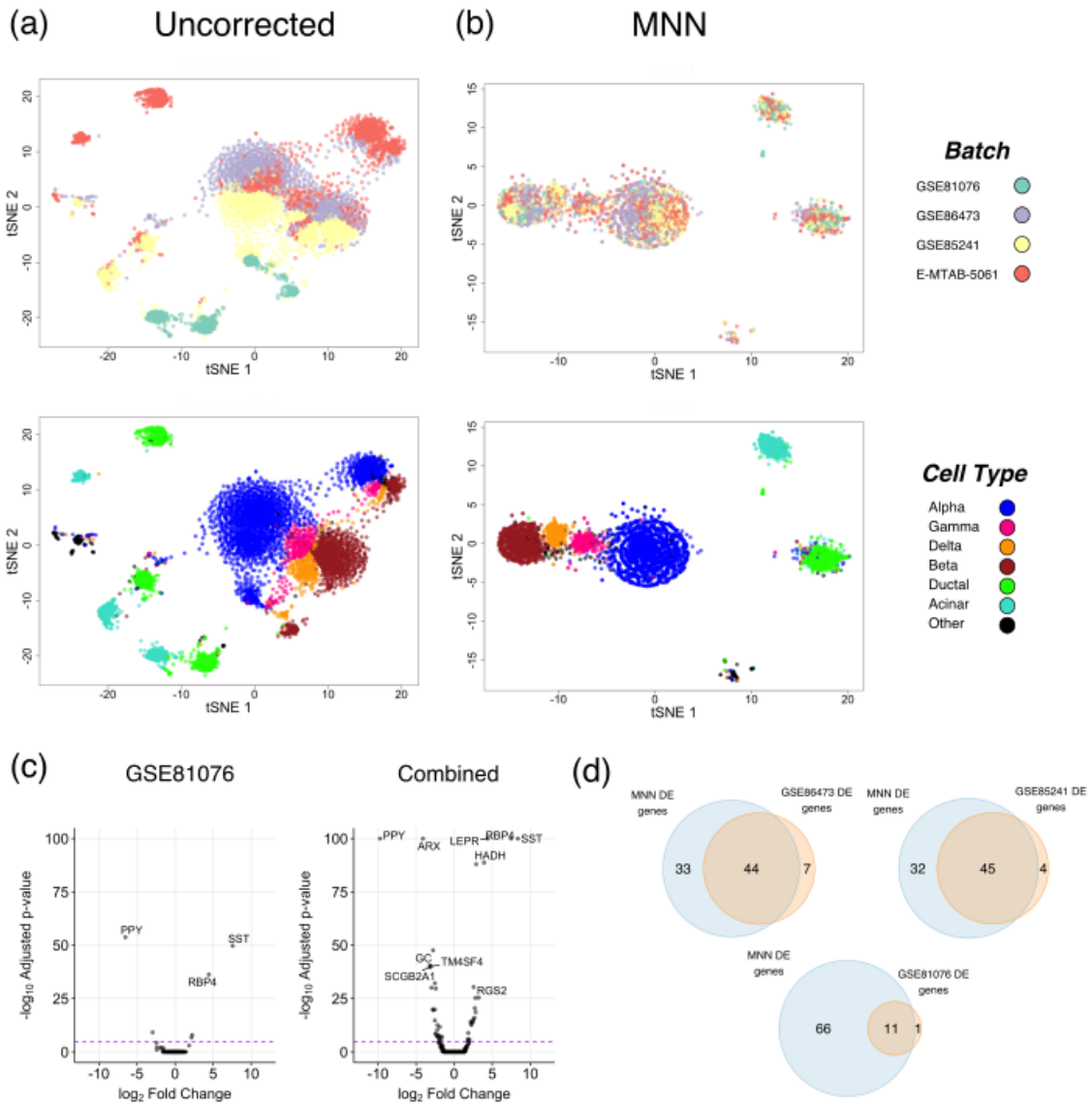

Figure 4: Application of MNN batch correction to pancreas cells using four data sets (GSE81076 with n=1007, GSE86473 with n= 2331, GSE85241 with n=1595 and E-MTAB-5061 with n=2163 cells) measured on two different platforms, CEL-seq(2) and SMART-seq2. *t*-SNE plots for (a) uncorrected (raw) data and (b) data corrected with our MNN method. The different batches are represented by four colours in the top panel of (a) and (b), whilst the different cell types are denoted in the bottom

panels by distinct colours. (c) Combining data sets by using MNN correction increases the power to detect differentially expressed genes. Volcano plots of differential expression testing in a single data set (GSE81076; δ-cells=54, γ-cells=19, left panel) and using the new cell type labels after MNN correction (Combined; δ-cells=428, γ-cells=425, right panel). The y-axis represents the $-\log_{10}$ Benjamini- Hochberg adjusted p-value ($-\log_{10}$ p-value > 100 are censored at 100 for comparable scales), and the x-axis is the $\log_2$ fold change of expression in cells over cells. Individual gene symbols are labelled where $|\log_2$ fold change$| > 3$. More genes are consistently differentially expressed at a FDR 5% in the combined data sets. (d) Venn diagrams representing the intersection of differentially expressed genes using the cell type labels after batch correction (blue circle) and using the original cell type labels from each individual study (orange circle). Numbers in each segment are the total number of DE genes between δ and γ islet cells in each batch. Each Venn diagram corresponds to a batch in which both cell types are present.

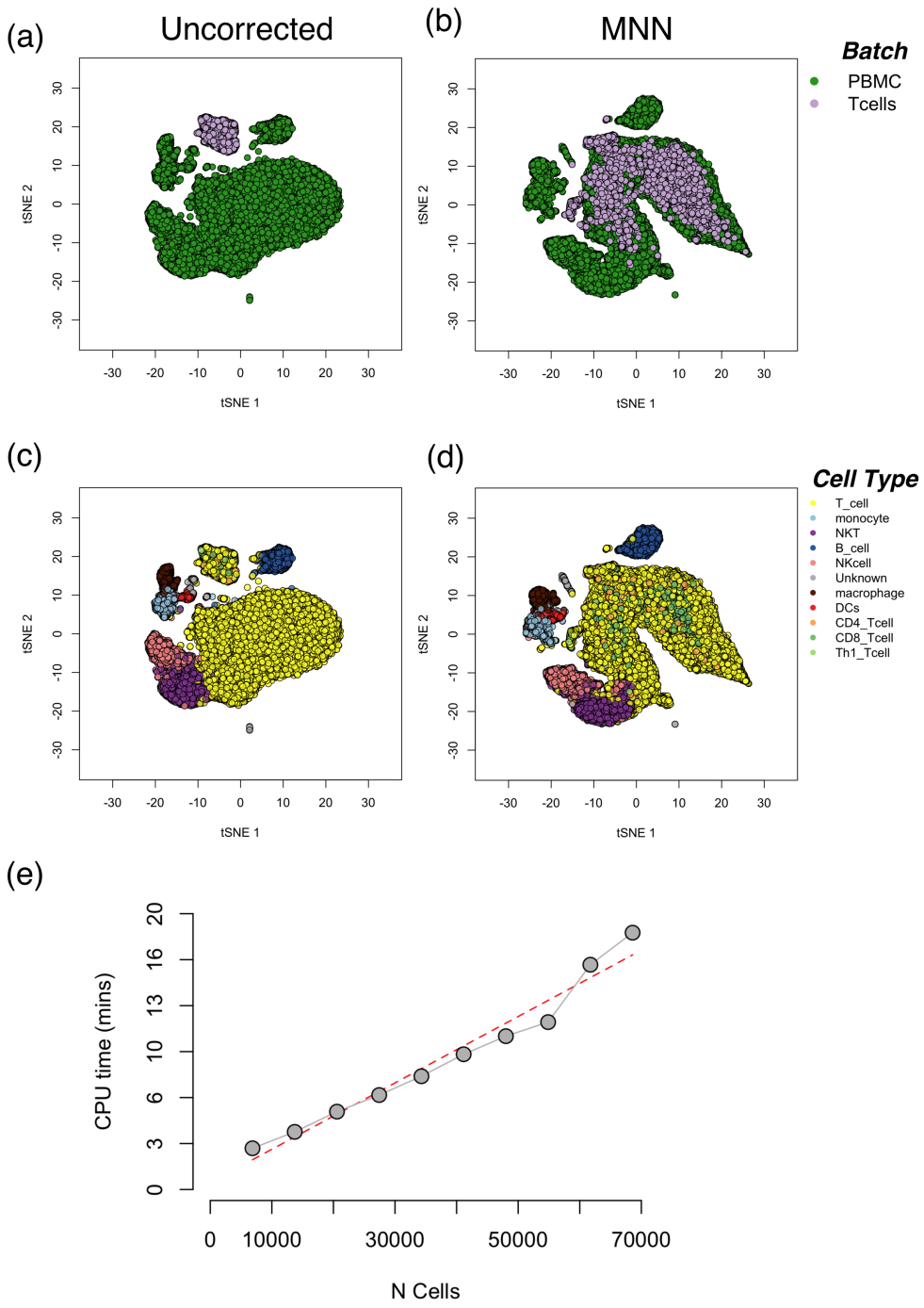

Figure 5: MNN batch correction scales to tens of thousands of cells. *t*-SNE plots of scRNA-seq data of human peripheral blood mononuclear cells and T cells (n=73039 cells), prior to batch correction (a, c) and following MNN correction (b, d). Individual points are coloured by their original cell type labels (c, d) and by the study batch of origin (a, b). (e) CPU time increases linearly in the number of input cells to MNN correction. Points represent the number of sub-sampled cells; the red dashed line represents the linear t between CPU time (minutes) and number of cells.

Discussion

Proper removal of batch effects is critical for valid data analysis and interpretation of the results. This is especially pertinent as the scale and scope of scRNA-seq experiments increase, exceeding the capacity of data generation within a single batch. To answer the relevant biological questions, merging data from different batches - generated by different protocols, operators and/or platforms - is required. However, for biological systems that are highly heterogeneous, it is likely that the composition of cell types and states will change across batches, due to stochastic and uncontrollable biological variability.

Existing batch correction methods do not account for differences in cell composition between batches and fail to fully remove the batch effect in such cases. This can lead to misleading conclusions whereby batch-specific clusters are incorrectly interpreted as distinct cell types. We demonstrate that our MNN method is able to successfully remove the batch effect in the presence of differences in composition, using both simulated data and real scRNA-seq data sets as well as demonstrating its scalability.

One prerequisite for our MNN method is that each batch contains at least one shared cell population with another batch. This is necessary for the correct identification of MNN pairs between batches. Batches without any shared structure are inherently difficult to correct, as the batch effects are completely confounded with biological differences. Such cases provide a motivation for using “cell controls”, i.e., an easily reproducible cell population of known composition (from a cell line for example) that is spiked into each sample for the purpose of removing batch effects across samples.

A notable feature of our MNN correction method is that it adjusts for local variations in the batch effects by using a Gaussian kernel. This means that our method can accommodate differences in the size or direction of the batch effect between different cell subpopulations in the high-dimensional space. Such differences are not

easily handled by methods based on linear models (as this would require explicit modelling of pre-defined groupings of cells, which would defeat the purpose of using scRNA-seq to study population heterogeneity in the first place). This also has some implications for the use of cell controls. Our results for the pancreas data set suggest that considering cell-type specific batch effects (the default setting of MNN) rather than a globally constant batch effect for all cells, improves batch removal results (Supplementary Figure 7). An important consequence is that a single cell control population might not suffice for accurate estimation of local batch effects. Rather, it may be necessary to use an appropriately mixed population of cells to properly account for local variation.

We have demonstrated in simulations and real data sets that MNN successfully combines cells with the same cell type label, by bringing cells from different batches onto a common coordinate system which is defined by the first (reference) batch, such that all batches can be analysed together. Therefore, MNN eliminates discrepancies between related batches without an analysis or interpretation of the origins and causes of batch effects (between each pair of batches). The study of technical and biological origins of these discrepancies may also be interesting. For instance, where one batch contains cells from a gene knock-out experiment and the other batch contains cells from a wild-type organism. In such cases we could potentially examine the correction vectors (provided as an output of the MNN algorithm) to understand the differences between batches.

Batch correction plays a critical role in the interpretation of data from scRNA-seq studies. This includes both small studies, where logistical constraints preclude the generation of data in a single batch; as well as those involving international consortia such as the Human Cell Atlas, where scRNA-seq data is generated on a variety of related tissues at different times and by multiple laboratories. Our MNN method provides a superior alternative to existing methods for batch correction in the presence of compositional differences between batches. We anticipate that it will improve the rigour of scRNA-seq data analysis and, thus, the quality of the biological conclusions.

Works Cited

1 Jaitin, Diego Adhemar and Kenigsberg, Ephraim and Keren-Shaul, Hadas and Elefant, Naama and Paul, Franziska and Zaretsky, Irina and Mildner, Alexander and Cohen, Nadav and Jung, Steffen and Tanay, Amos and others. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343, 6172 (2014), 776–779.

2 Klein, Allon M and Mazutis, Linas and Akartuna, Ilke and Tallapragada, Naren and Veres, Adrian and Li, Victor and Peshkin, Leonid and Weitz, David A and Kirschner, Marc W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161, 5 (2015), 1187-1201.

3 Macosko, Evan Z and Basu, Anindita and Satija, Rahul and Nemesh, James and Shekhar, Karthik and Goldman, Melissa and Tirosh, Itay and Bialas, Allison R and Kamitaki, Nolan and Martersteck, Emily M and others. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161, 5 (2015), 1202-1214.

4 Gierahn, Todd M and Wadsworth II, Marc H and Hughes, Travis K and Bryson, Bryan D and Butler, Andrew and Satija, Rahul and Fortune, Sarah and Love, J Christopher and Shalek, Alex K. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, 14, 4 (2017), 395–398.

5 Hicks, Stephanie C, F. William Townes, Mingxiang Teng, and Rafael A. Irizarry. Missing data and technical variability in single-cell RNA- sequencing experiments. *BioRxiv* (2017), https://doi.org/10.1101/025528.

6 Tung, Po-Yuan and Blischak, John D and Hsiao, Chiaowen Joyce and Knowles, David A and Burnett, Jonathan E and Pritchard, Jonathan K and Gilad, Yoav. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7 (2017), 39921.

7 Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43, 7 (2015), e47–e47.

8 Johnson, W Evan and Li, Cheng and Rabinovic, Ariel. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8, 1 (2007), 118-127.

9 Risso, Davide and Ngai, John and Speed, Terence P and Dudoit, Sandrine. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnoly*, 32, 9 (2014), 896-902.

10 Leek, Jeffrey T. Svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42, 21 (2014).

11 Spitzer, Matthew H, Pier Federico Gherardini, Gabriela K Fragiadakis, Nupur Bhattacharya, Robert T Yuan, Andrew N Hotson, Rachel Finck, et al. An interactive reference framework for modeling a dynamic immune system. *Science*, 349, 6244

(2015), 1259425.

12 Nestorowa, Sonia, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128, 8 (2016), e20–e31.

13 Scialdone, Antonio, Yosuke Tanaka, Wajid Jawaid, Victoria Moignard, Nicola K Wilson, Iain C Macaulay, John C Marioni, and Berthold Göttgens. Resolving early mesoderm diversification through single-cell expression profiling. *Nature*, 535, 7611 (2016), 289--293.

14 Baron, Maayan, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3, 4 (2016), 346–360.

15 Bendall, Sean C, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human B Cell development. *Cell*, 157, 3 (2014), 714–725.

16 Maaten, Laurens van der, and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (2008), 2579–2605.

17 Picelli, Simone, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10, 11 (2013), 1096–98.

18 Paul, Franziska, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163, 7 (2015), 1663–1677.

19 Angerer, Philipp, Laleh Haghverdi, Maren Büttner, Fabian J Theis, Carsten Marr, and Florian Buettner. Destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 32, 8 (2015), 1241–1243.

20 Grün, Dominic, MauroJ Muraro, Jean-Charles Boisset, Kay Wiebrands, Anna Lyubimova, Gitanjali Dharmadhikari, Maaike vandenBorn, et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell*, 19, 2 (2016), 266-77.

21 Muraro, MauroJ, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon vanGurp, et al. A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3, 4 (2016), 385–394.e3.

22 Lawlor, Nathan, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V. Sivakamasundari, Ina Kycia, Paul Robson, and Michael L. Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Research*, 27, 2 (2017), 208–22.

23 Segerstolpe, Åsa, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andréasson, Xiaoyan Sun, Simone Picelli, et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabolism*, 24, 4 (2016), 593–607.

24 Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan

Wilson, Solongo B. Ziraldo, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8 (2017), 14049.

Online Methods

Generation and analysis of simulated data

We considered a three-component Gaussian mixture model in two dimensions (to represent the low dimensional biological subspace), where each mixture component represents a different simulated cell type. Two data sets with N = 1000 cells were drawn with different mixing coefficients (0.2, 0.3 and 0.5 for the first batch and 0.05, 0.65 and 0.3 for the second batch) for the three cell types. We then projected both data sets to G = 100 dimensions using the same random Gaussian matrix, thus simulating high-dimensional gene expression. Batch effects were incorporated by generating a Gaussian random vector for each data set and adding it to the expression profiles for all cells in that data set.

Processing and analysis of the haemaopoetic data sets

Gene expression counts generated by Nestorowa *et al.* [12] on the SMART-seq2 platform (1920 cells in total) were downloaded from the NCBI Gene Expression Omnibus (GEO) using the accession number GSE81682. Expression counts generated by Paul *et al*. [18] on the MARS-seq platform (10368 cells in total) were obtained from NCBI GEO using the accession GSE72857. Then, using FACS sorting, the authors identified 2729 myeloid progenitor cells (CMP, GMP, and MEP) as Lin– (lineage negative) c-Kit+ Sca1– and gated the cells further by the levels of the FcgR and CD34 markers; these cells were used for the analysis in this manuscript. For batch correction, we identified a set of 3937 common highly variable genes between the two data sets, by applying the method described by Brennecke *et al.* [25] to each data set. For both data sets, we performed library size normalization before log-transforming the normalized expression values. A priori cell labels were assigned to each cell based on the original publications.

Processing and analysis of the pancreas data sets

Raw data were obtained from NCBI GEO using the accession numbers GSE81076 [20] (CEL-seq), GSE85241 [21](CEL-seq2) and GSE86473 [22] (SMART-seq2); or from ArrayExpress, using the accession E-MTAB-5061 [23] (SMART-seq2). Count matrices were used as provided by GEO or ArrayExpress, if available. For GSE86473, reads were aligned to the hg38 build of the human genome using STAR version 2.4.2a [26] with default parameters, and assigned to Ensembl build 86 protein-coding genes using featureCounts version 1.4.6 [27].

Quality control was performed on each data set independently to remove poor quality cells (>20% of total counts from spike-in transcripts, <100,000 reads, >40% total counts from ribosomal RNA genes). Sparse cells and genes (90% zero values) were also removed, leaving a total of 7236 cells available across all 4 data sets. Normalization of cell-specific biases was performed for each data set using the deconvolution method of Lun *et al.* [28]. Counts were divided by size factors to obtain normalised expression values that were log-transformed after adding a pseudo-count of 1. Highly variable genes were identified in each data set using the method of Brennecke *at al.* [25]. We took the union of highly variable genes that are commonly expressed across all four data sets, resulting in 2507 genes that were used for the MNN batch correction.

Cell type labels for each data set were assigned based on the provided metadata (GSE86473, EMTAB-5061) or, if the labels were not provided, were inferred from the data using the method employed in the original publication (GSE81076, GSE85241).

To demonstrate the utility of our batch correction method in downstream analyses, we applied dimensionality reduction (*t*-SNE) to the MNN-corrected expression matrix from the pooled pancreas data sets. We constructed a shared-nearest-neighbour (SNN) graph [29] using the combined cells and the union of the highly variable genes that were commonly expressed across all data set. To identify communities of cells

we applied the “Walktrap" algorithm to the SNN graph [30], with 5 steps. This identified a total of 11 clusters. To assign specific cell type labels to these clusters, we examined the expression of the marker genes that were used for cell type assignment in the original publications. Specifically, *GCG* was used to mark α-islets, *INS* for β-islets, *SST* for δ-islets, *PPY* for γ-islets, *PRSS1* for acinar cells, *KRT19* for ductal cells, and *COL1A1* for mesenchyme cells. Cells in the cluster with the highest expression of each marker gene were assigned to the corresponding cell type. All remaining cells were allocated into an additional “Unassigned/Unknown" cluster.

The differential expression analysis was performed using methods from the limma package [7]. For the analysis on all cells, we parameterized the design matrix such that each batch-cluster combination formed a separate group in a one-way layout using the labels derived from the batch-corrected data (see above). We used this design to fit a linear model to the normalized uncorrected log-expression values for each gene, and performed empirical Bayes shrinkage to stabilize the sample variances. A moderated t-test was applied to compare the δ- and γ-islet clusters across all batches. Specifically, we tested whether the average expression of each cluster across all batches was equal between the two cell types. Differentially expressed genes were defined as those detected at an FDR of 5%. For comparison, we repeated this analysis for each batch using only cells from batches where both cell types were present. Here, we used a design matrix with a one-way layout constructed from the original cell type assignments. δ- and γ-islet cell types were directly compared within this batch.

Application of batch correction to droplet-based data

Single-cell gene expression measurements derived from the droplet-based platform by 10X Genomics using their Chromium v2 chemistry were downloaded from the company website (https://support.10xgenomics.com/single-cell-gene-expression/datasets). Expression data from 4459 human T cells (t_4k) and 68,580 peripheral blood mononuclear cells (PBMCs; pbmc68k) from two separate donors were normalised

separately using size factors estimated by the deconvolution method as previously described [28]. Highly variable genes were defined within each data set as previously described [25] (PBMC - 1409 genes, T cells - 1219). To define communities of transcriptionally similar cells, we constructed a SNN graph, and assigned cells to specific communities using the Walktrap algorithm. The identity of each community was assigned by visualisation of canonical marker gene expression to major leukocyte lineages (CD3, CD20, CD14, CD16, CD1C, CD56). Droplet data sets were combined using our MNN approach on the intersection of the two highly variable gene sets (270 genes). Low-dimensional representations of individual and combined data sets was performed using *t*-SNE.

MNN correction scalability

Scalability testing of our MNN correction method was performed by random sampling of cells between 10 and 100% of the total number of PBMCs, i.e., where 100% = 68,000 cells. We combined each subset with the set of 4459 T cells, and recorded the CPU time in the R environment (R Core Team 2017) using the *system.time* function. For each combination of data, the R environment garbage collector was invoked *prior* to recording the function call system time.

t-SNE plots

We generated the *t*-SNE plots using the Rtsne package with identical parameter settings for the uncorrected and batch corrected data using MNN, limma and ComBat. In all plots, we have used the distance matrix as the input for the Rtsne function (i.e., Rtsne parameter *is.distance=TRUE*). For the haematopoietic data where continuity of data structure is expected, we accounted for this by choosing a large perplexity parameter (i.e., 90). For all other data sets where existence of separate clusters in the data is expected, we have used the default perplexity parameter (i.e., 30), and again have used identical parameter settings across all batch correction methods.

Silhouette coefficient

To assess the separation of the cell types for the pancreas data, we computed the silhouette coefficient using the kBET package in R [31]. Here, each unique cell type label defines a cluster of cells. Let $a(i)$ be the average distance of cell i to all other cells within the same cluster as i, and $b(i)$ be the average distance of cell i to all cells assigned to the neighbouring cluster, i.e., the cluster with the lowest average distance to the cluster of i. The Silhouette coefficient for cell i is defined as:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & if\ a(i) > b(i) \end{cases} \qquad (2)$$

A larger $s(i)$ implies that the cluster assignment for cell i is appropriate, i.e., it is close to other cells in the same cluster yet distant from cells in other clusters. As dimensionality reduction by *t*-SNE facilitates more reasonable clustering results compared to clustering in the high dimensions, we calculated the silhouette coefficients using distance matrices computed from the *t*-SNE coordinates of each cell in the batch-corrected and the uncorrected data.

Entropy of batch mixing

Entropy of mixing [32] for c different batches is defined as:

$$E = \sum_{i=1}^{c} x_i \log(x_i) \qquad (3)$$

where x_i is the proportion of cells from batch i in a given region, such that $\sum_{i=1}^{c} x_i = 1$. We assessed the total entropy of batch mixing on the first two PCs of the batch-corrected and the uncorrected pancreas data sets, using regional mixing entropies according to Equation 3 at the location of 100 randomly chosen cells from all batches. The regional proportion of cells from each batch was defined from the set of 100 nearest neighbours for each randomly chosen cell. The total mixing entropy was then calculated as the sum of the regional entropies. We repeated this for 100 iterations with different randomly chosen cells to generate boxplots of the total entropy (Supplementary Figures 5q and 6q).

A Life Sciences Reporting Summary is available.

Works Cited

25 Brennecke, Philip, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, et al. Accounting for technical noise in single-sell RNA-seq experiments. *Nature Methods*, 10, 11 (2013), 1093-95.

26 Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 1 (2013), 15-21.

27 Liao, Yang and Smyth, Gordon K and Shi, Wei. FeatureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30, 7 (2014), 923-30.

28 Lun, Aaron TL and Bach, Karsten and Marioni, John C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17, 1 (2016), 75.

29 Xu, Chen, and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering Method. *Bioinformatics*, 31, 12 (2015), 1974-80.

30 Pons, Pascal and Latapy, Matthieu. Computing communities in large networks using random walks. *ISCIS*, 3733 (2005), 284--293.

31 Buttner, Maren and Miao, Zhichao and Wolf, Alexander and Teichmann, Sarah A and Theis, Fabian J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. *BioRxiv, doi: https://doi.org/10.1101/200345* (2017), 200345.

32 Brandani, Giovanni B, Marieke Schor, Cait E MacPhee, Helmut Grubmüller, Ulrich Zachariae, and Davide Marenduzzo. Quantifying disorder through conditional entropy: an application to fluid mixing. *PloS One*, 8, 6 (2013), e65617.

Data availability

The published data sets used in this manuscript are available through GEO accession numbers
SMART-seq2 platform haematopoietic data by Nestorowa *et al.* [12]: GSE81682,
MARS-seq platform haematopoietic data by Paul *et al.* [18] : GSE72857,
CEL-seq platform pancreas data by Grün *et al.* [20]: GSE81076,
CEL-seq2 platform pancreas data by Muraro *et al.* [21]: GSE85241,
SMART-seq2 platform pancreas data by Lawlor *et al.* [22]: GSE86473,
or ArrayExpress accession number:
SMART-seq2 platform pancreas data by Segerstolpe *et al.* [23]: E-MTAB-5061.

Software availability

An open-source software implementation of our MNN method is available as the *mnnCorrect* function in version 1.6.2 of the scran package on Bioconductor (https://bioconductor.org/packages/scran). All code for producing results and figures in this manuscript are available on Github (https://github.com/MarioniLab/MNN2017).

Acknowledgements

We are grateful to Fiona K. Hamey, James P. Munro, Jonathan Griffiths and Maren Büttner for helpful discussions. LH was supported by Wellcome Trust Grant 108437/Z/15 to JCM. ATLL was supported by core funding from CRUK (award number 17197 to JCM). MDM was supported by Wellcome Trust Grant 105045/Z/14/Z to JCM. JCM was supported by core funding from EMBL and from CRUK (award number 17197).

Author contributions

L.H. developed the method and the computational tools, performed the analysis, and wrote the paper. A.T.L.L. developed the method and the computational tools and wrote the paper. M.D.M. developed the method, performed the analysis and wrote the paper. J.C.M . developed the method, wrote the paper and supervised the study.

Competing interests

The authors declare no competing financial interests.