nature biotechnology

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

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

Large-scale single-cell RNA sequencing (scRNA-seq) data sets that are produced in different laboratories and at different times contain batch effects that may compromise the integration and interpretation of the data. Existing scRNA-seq analysis methods incorrectly assume that the composition of cell populations is either known or identical across batches. We present a strategy for batch correction based on the detection of mutual nearest neighbors (MNNs) in the high-dimensional expression space. Our approach does not rely on predefined or equal population compositions across batches; instead, it requires only that a subset of the population be shared between batches. We demonstrate the superiority of our approach compared with existing methods by 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 can be scaled to large numbers of cells.

The decreasing cost of scRNA-seq experiments[1–4] 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 be 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], in which the expression of genes in one batch differs systematically from that in another batch. Such differences can mask underlying biology or introduce spurious structure in the data; thus, to avoid misleading conclusions, they must be corrected before further analysis.

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, thus 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 procedure 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 their focus is primarily on identifying unknown factors of variation, for example, those due to unrecorded experimental differences in cell processing. After 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 application to scRNA-seq data is based on the assumption that the composition of the cell population within each batch is identical. Any systematic differences in mean gene expression between batches are attributed to technical differences that can be regressed out. However, in practice, the population composition is usually not identical across batches in scRNA-seq studies. Even if 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 procedures such as cell culture or tissue extraction, dissociation and sorting. Consequently, the estimated coefficients for the batch blocking factors are not purely technical but contain a nonzero biological component because of differences in composition. Batch correction based on these coefficients would thus yield inaccurate representations of the cellular expression profiles, and the results might potentially be worse than if no correction were 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 for this approach is that a given cell type in the reference batch will be most similar to cells of its own type in the new batch. Such projection strategies can be applied by using several dimensionality-reduction methods, such as principal component analysis (PCA) or diffusion maps, or by using force-based methods such as *t*-distributed stochastic neighbor 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**).

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [2]Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany. [3]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [4]Wellcome Trust Sanger Institute, Cambridge, UK. Correspondence should be addressed to J.C.M. (john.marioni@cruk.cam.ac.uk).

Received 3 July 2017; accepted 1 February 2018; published online 2 April 2018; doi:10.1038/nbt.4091

Here, we propose a new method for removal of discrepancies between biologically related batches according to the presence of 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 an 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 neighbors for batch correction

Our approach identifies cells that have mutually similar expression profiles between different experimental batches or replicates. 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 artifacts) 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 (for example, intersample differences due to genotype). After correction, multiple batches can be 'joined up' into a single data set (**Fig. 1**).

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

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

Subsequently, we compute the Euclidean distance between the cosine-normalized expression profiles of pairs of cells. Calculating Euclidean distances on these normalized data is equivalent to using cosine distances on the original expression values (**Supplementary Note 2**). Cosine distances have been widely used for measuring cell similarities according to expression profiles[11,13–15] and are appealing because they are scale independent[15] and thus robust to technical differences in sequencing depth and capture efficiency between batches.

The next step involves identification of mutual nearest neighbors. Consider an scRNA-seq experiment consisting of two batches 1 and 2. For each cell i_1 in batch 1, we find the k cells in batch 2 with the smallest distances to i_1, i.e., its k nearest neighbors in batch 2. We do the same for each cell in batch 2 to find its k nearest neighbors in batch 1. If a pair of cells from each batch is contained in each other's set of nearest neighbors, those cells are considered to be mutual nearest neighbors (**Fig. 1**). We interpret these pairs as containing cells that belong to the same cell type or state despite being generated in different batches. Thus, 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) the batch-effect variation is much smaller than the biological-effect variation between different cell types (more detailed discussion of these assumptions in **Supplementary Note 3**). The biological subspace refers to a set of basis vectors that represent biological processes; the length of each vector is equal to the number of genes. For example, some of these vectors may represent the cell cycle; some vectors may define expression profiles specific to each cell type; and other vectors may represent differentiation or activation states. The true expression profile of each cell can be expressed as the 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 assumption can be more intuitively understood in that cells from the same population in different batches form parallel hyperplanes with respect to each other (**Fig. 1**). 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. Although a set of biologically relevant genes (for example, 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

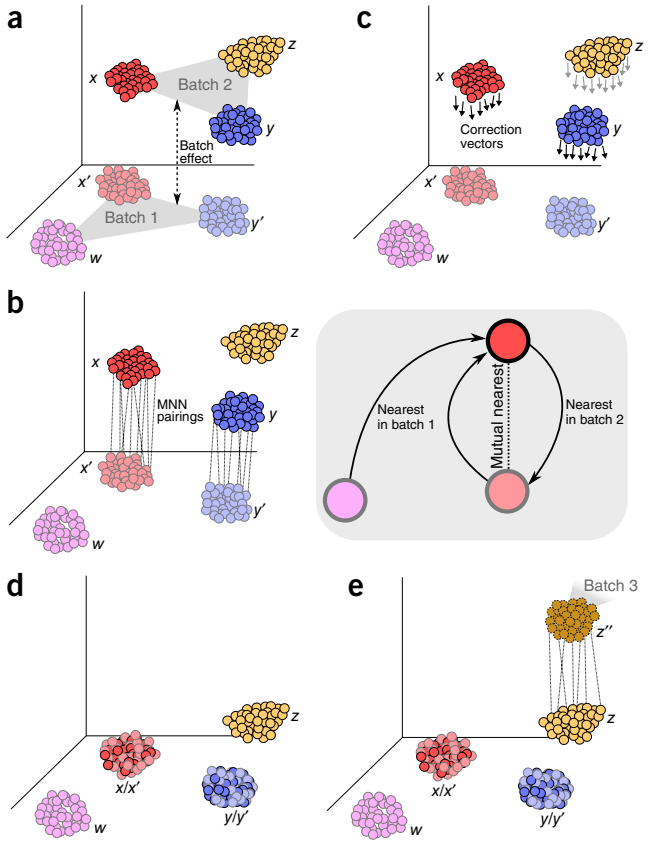

Figure 1 Schematics of batch-effect correction by MNN. (**a**) Batch 1 and batch 2 in high dimensions with an almost orthogonal batch effect difference between them. (**b**) The algorithm identifies matching cell types by finding MNN pairs of cells (gray box). (**c**) Batch-correction vectors are calculated 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 the reference, and the procedure is repeated for integration of any new batch.

batch-correction vector is then calculated as a weighted average of these pair-specific vectors, as computed with 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 contributes to the batch effect for cells in the neighborhood 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 Fig. 1**). We emphasize that this correction is performed for all cells, regardless of whether they participate in a MNN pair. Thus, correction can be performed on all cells in each batch, even if they do not have a corresponding cell type in the other batches.

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 of three batch-correction methods—our MNN-based correction method, limma and ComBat—to the simulated data, then evaluated the results by inspecting *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, such that each cluster contains a mixture of cells from both batches. However, we observed this ideal result only after MNN correction (**Fig. 2**). Expression data that were uncorrected or corrected with the other methods exhibited at least one cluster containing cells from only a single batch, thus indicating that the batch effect was not fully removed. This result 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 Fig. 2**).

MNN correction outperforms existing methods on hematopoietic data

To demonstrate the applicability of our method to real data, we considered two hematopoietic data sets generated in different laboratories through two different scRNA-seq protocols. In the first data set[12], the authors used the SMART-seq2 protocol[17] to profile single cells from hematopoietic stem and progenitor cell populations in 12-week-old female mice. Using marker expression profiles from fluorescence-activated cell sorting (FACS), we retrospectively assigned known cell-type labels to cells (Online Methods). These labels included multipotent progenitors, lymphoid-primed multipotent progenitors, hematopoietic stem and progenitor cells, hematopoietic stem cells, common myeloid progenitors (CMPs), granulocyte–monocyte progenitors (GMPs) and megakaryocyte–erythrocyte progenitors (MEPs). In the second data set[18], the authors used the massively parallel single-cell RNA-sequencing (MARS-seq) protocol to assess single-cell heterogeneity in myeloid progenitors from 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 for the highly variable genes, before and after correction with each of the three methods (MNN, limma and ComBat) (**Fig. 3a–d** and Online Methods). Only MNN correction correctly merged the cell types that were shared between batches, i.e., CMPs, MEPs and GMPs, while preserving the underlying differentiation hierarchy[12,18] (**Fig. 3e**). In contrast, the shared cell types still clustered by batch after correction with limma or ComBat, thus indicating that the batch effect had not been completely removed (coloring by batch in **Supplementary Fig. 3**). This result is attributable to the differences in cell-type composition between batches and is 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 cell types in common between the two batches (**Fig. 3f–i**). Among the methods, MNN correction was still the most effective at removing the batch effect.

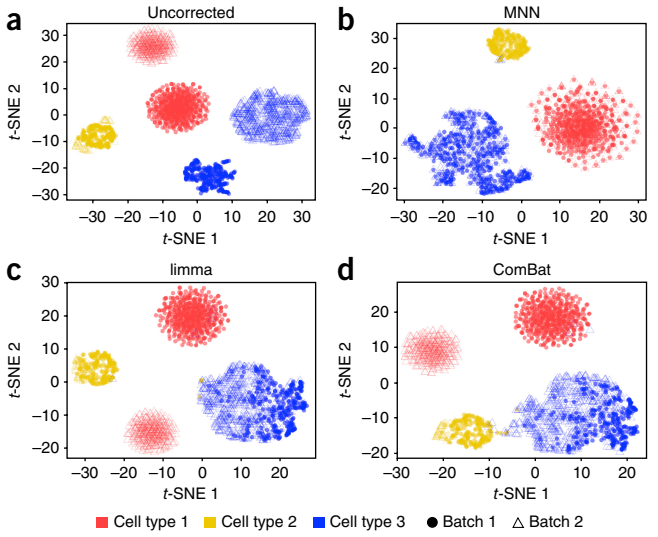

Figure 2 *t*-SNE plots of simulated scRNA-seq data containing two batches of different cell types (with each batch containing $n = 1{,}000$ cells). (**a**–**d**) Data before correction (**a**) and after correction with our MNN method (**b**), limma (**c**) or ComBat (**d**). In this simulation, each batch (closed circle or open triangle) contained different numbers of cells in each of three cell types (specified by color).

As a justification for the orthogonality of the 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 components of the reference batch used in MNN (i.e., the SMART-seq2 data set). Most angles are close to 90°, thus supporting the near-orthogonality assumption (**Supplementary Fig. 3e**). A diffusion map[19] of the MNN-corrected data (**Supplementary Fig. 3f–h**) shows the same differentiation hierarchy of cell types as that in **Figure 3e**. Repeating the same analysis on a subset of randomly sampled genes (1,500 out of the total of 3,904 highly variable genes) yielded similar results, thus demonstrating the robustness of our analysis with respect to the input gene set (**Supplementary Fig. 4**).

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 through a variety of methods. Here, we focused on the pancreas because 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[20–23] generated through two different scRNA-seq protocols (SMART-seq2 and scRNA-seq by multiplexed linear amplification (CEL-seq)/CEL-seq2). Cell-type labels were taken from the provided metadata or were derived according to the methodology described in the original publication (further details of data preprocessing in Online Methods).

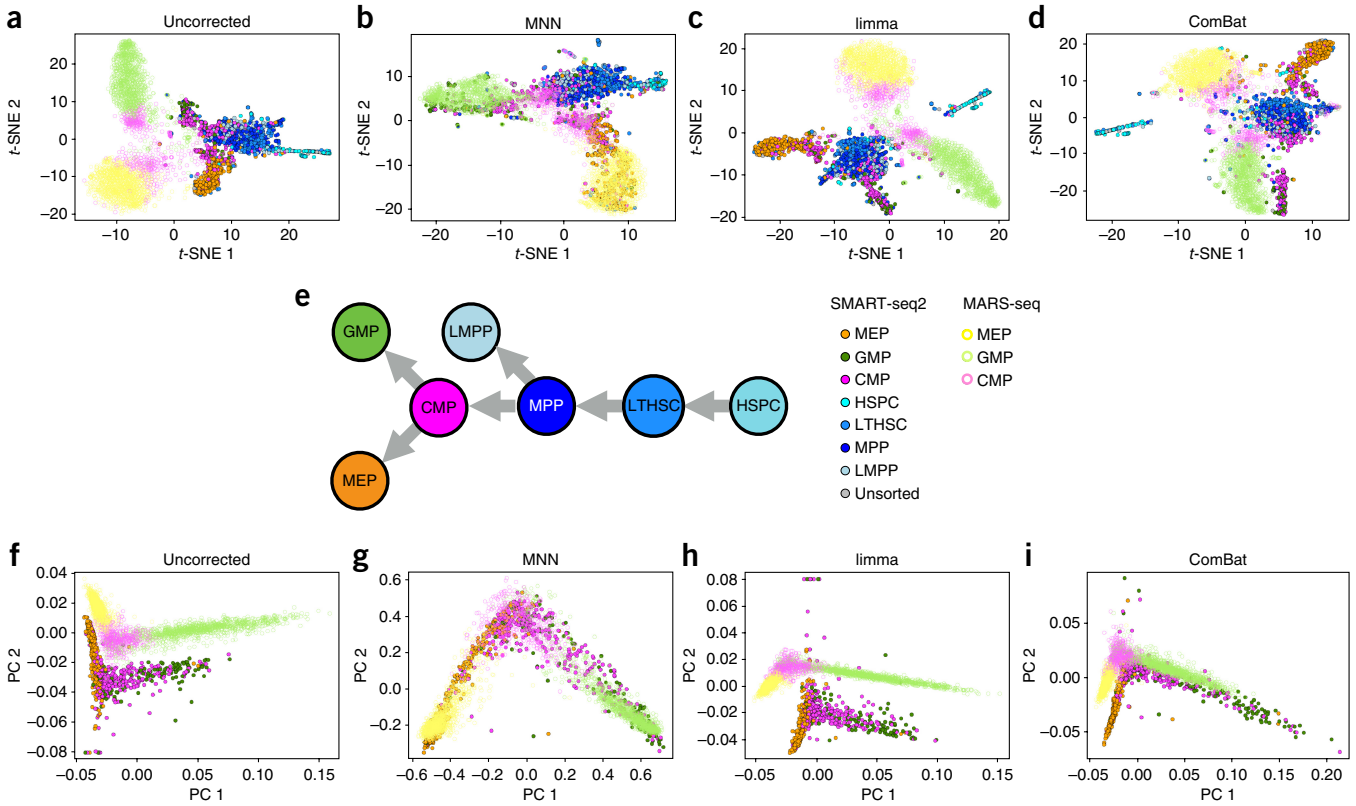

Figure 3 *t*-SNE plots of scRNA-seq count data for cells from the hematopoietic lineage, prepared in two batches by using different technologies (SMART-seq2 with $n = 1{,}920$ cells, closed circles; MARS-seq with $n = 2{,}729$ cells, open circles). (**a**–**d**) Plots generated before batch correction (**a**) and after batch correction with our MNN method (**b**), limma (**c**) or ComBat (**d**). Cells are colored according to their annotated cell type. (**e**) The expected hierarchy of hematopoietic cell types. (**f**–**i**) PCA plots of scRNA-seq count data for common cell types between the two batches of the hematopoietic lineage generated (SMART-seq2 with $n = 791$ cells; MARS-seq with $n = 2{,}729$ cells) before batch correction (**f**) and after batch correction through our MNN method (**g**), limma (**h**) or ComBat (**i**). MPP, multipotent progenitors; LMPP, lymphoid-primed multipotent progenitors; LTHSC, long-term hematopoietic stem cells; HSPC, hematopoietic stem and progenitor cells; PC, principal component.

We applied MNN, limma and ComBat to the combined data set and examined the corrected data. All three batch-correction methods improved the grouping of cells by their cell-type labels (**Fig. 4a**,**b**, Online Methods and **Supplementary Fig. 5a**–**d**). This result is not surprising, because the discrepancy between cell-type composition in the four batches was modest (**Supplementary Table 1**). However, even a small difference in composition was sufficient to cause ductal and acinar cells to be incorrectly separated after correction with limma or ComBat. By comparison, both cell types were coherently grouped across batches after MNN correction, in agreement 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 Fig. 5** and Online Methods). The average silhouette coefficient after MNN correction was significantly larger than those in the uncorrected and limma- and ComBat-corrected data ($P < 0{:}05$, two-sided Welch's *t*-test). Thus, MNN correction is able to decrease the between-batch variance within each cell type while preserving differences among cell types. We also computed the entropy of mixing (Online Methods) to quantify the extent of intermingling of cells from different batches. The data that were batch corrected with MNN showed higher entropy of mixing than did the uncorrected data and the data corrected with limma or ComBat (**Supplementary Fig. 5**). The improvement in the mixing of batches was observed in the reduced-dimension space obtained through either *t*-SNE or PCA (**Supplementary Fig. 5e**–**l**). We again supported our assumption that batch effects are adequately removed when they lie orthogonally to the biological subspace (**Supplementary Fig. 5m**–**o**). The observed structure in the pancreas data was robust to the size of the input gene set, as demonstrated by random subsampling of the total highly variable gene set (**Supplementary Fig. 6**).

MNN correction improves differential expression analyses

After batch correction is performed, the corrected expression values can be used in routine downstream analyses such as clustering prior to differential gene expression identification. To provide a demonstration, 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, whereas 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 procedure ensured that the cluster labels were directly comparable between cells in different batches. In contrast, if clustering had been performed separately in each batch, there would have been no guarantee that a (weakly separated) cluster detected in one batch would have had a direct counterpart in another batch.

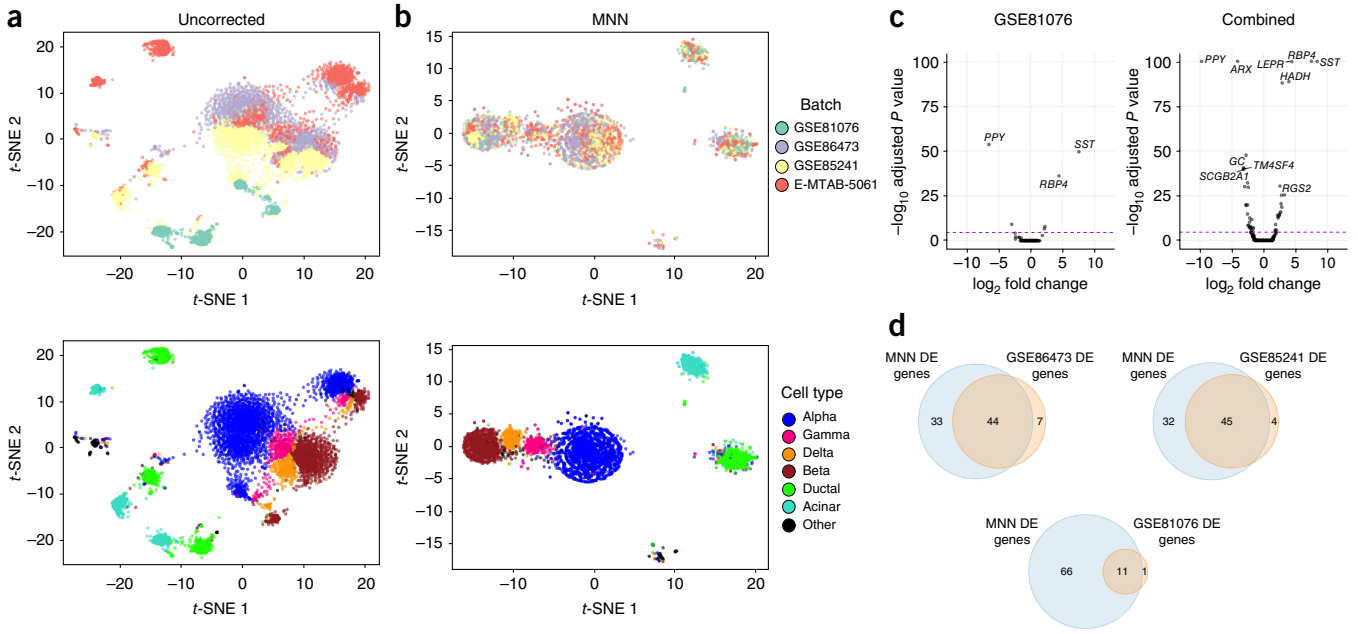

Figure 4 Application of MNN batch correction to pancreas cells by using four data sets (GSE81076 with n = 1,007 cells, GSE86473 with n = 2,331 cells, GSE85241 with n = 1,595 cells and E-MTAB-5061 with n = 2,163 cells) measured on two different platforms, CEL-seq2 and SMART-seq2. (**a**,**b**) t-SNE plots for uncorrected (raw) data (**a**) and data corrected with our MNN method (**b**). At top, the different batches are represented by four colors; at bottom, the different cell types are denoted by distinct colors. (**c**) Combining data sets through MNN correction increases the power to detect DE genes. Volcano plots of differential expression testing in a single data set (GSE81076; delta cells = 54, gamma cells = 19, left) and by using the new cell-type labels after MNN correction (combined; delta cells = 428, gamma cells = 425, right). The y axis represents the $-\log_{10}$ Benjamini–Hochberg-adjusted P value ($-\log_{10}P$ values >100 are censored at 100 for comparable scales), and the x axis is the $\log_2$ fold change of expression in gamma cells over delta cells. Individual gene symbols are labeled when $|\log_2$ fold change$| > 3$. More genes are consistently differentially expressed at an FDR of 5% in the combined data sets. (**d**) Venn diagrams representing the intersection of DE genes by using the cell-type labels after batch correction (blue circles) and using the original cell-type labels from each individual study (orange circles). Numbers in each segment are the total numbers of DE genes between delta and gamma islet cells in each batch. Each Venn diagram corresponds to a batch in which both cell types are present.

We used our new clusters to perform a differential expression analysis between the delta islet cluster and the gamma islet cluster. Using cells from all batches, we detected 76 differentially expressed (DE) genes at a false discovery rate (FDR) of 5% (**Fig. 4c**). This set included the marker genes for the cells included in the analysis (*PPY* and *SST*), genes involved in pancreatic islet cell development (*PAX6*) and genes recently implicated in delta islet function and the development of type 2 diabetes (*CD9* and *HADH*)[22]. For comparison, we repeated the differential expression analysis by using only cells from each batch in which both cell types were present[20–22]. The results yielded only 12, 49 and 51 genes, respectively, at an FDR of 5%, which encompassed 14.5–57.9% of those detected when all cells were used (**Fig. 4d**). Merging data sets is beneficial because it increases the number of cells without requiring additional experimental work, improves statistical power for downstream analyses such as differential gene expression and consequently provides additional biological insights. To this end, our MNN approach is critical because 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 for 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 comprising 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, whereas peripheral T cells contain a mixture of naive and antigen-exposed lymphocytes involved in active immune surveillance.

A naive merging of these two data sets without accounting for batch effects illustrated the separation of the T cells from their counterparts in the PBMC data (**Fig. 5a**,**b**). The combination of these two data sets by using MNNs demonstrated that the separate peripheral T cells mapped to the T cell subsets within the PBMC mixture (**Fig. 5c**,**d**). Importantly, other peripheral lymphocyte relationships were not distorted by the correction applied, despite the absence of MNNs in the T cell data set (**Fig. 5c**). Specifically, 4,446/4,459 (99.7%) of individual T cells mapped onto their appropriate counterparts in the PBMC data set (**Fig. 5**). The remaining 13/4,459 (0.3%) mapped primarily to a small cluster of unknown ontogeny and to the edges of a large cluster of monocytes. In contrast, 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, then corrected the batch effect between each subsample and the 4K T cell data. Within the range of 7,000 to 70,000 cells, we observed

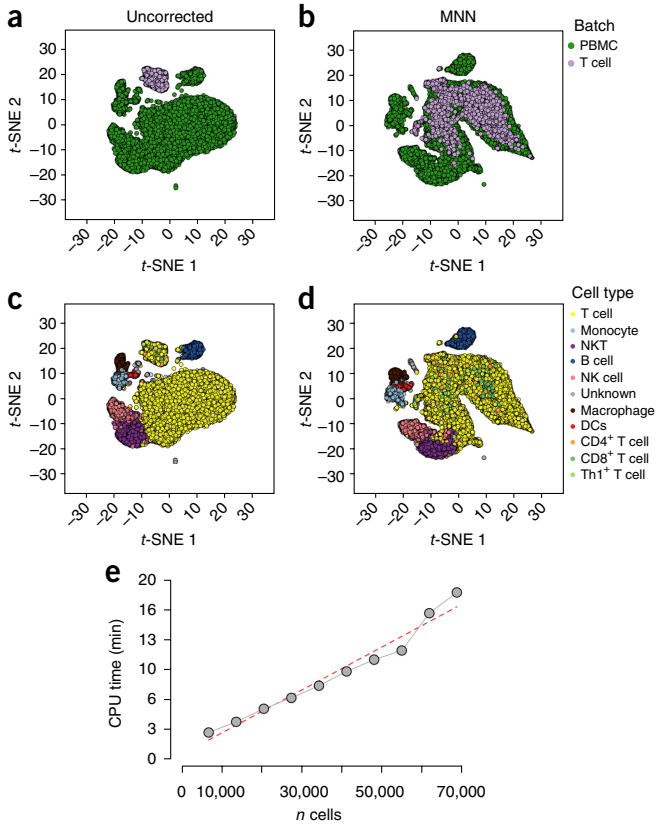

Figure 5 MNN batch correction can be scaled to tens of thousands of cells. (**a**–**d**) *t*-SNE plots of scRNA-seq data of human PBMCs and T cells ($n = 73{,}039$ cells), before batch correction (**a**,**c**) and after MNN correction (**b**,**d**). Individual points are colored according to their original cell-type labels (**c**,**d**) and the study batch of origin (**a**,**b**). (**e**) Central processing unit (CPU) time increases linearly with the number of cells input to MNN correction. Points represent the number of subsampled cells; the red dashed line represents the linear relationship between CPU time (minutes) and number of cells.

an approximately linear time increase (**Fig. 5e**). Thus, our method is compatible with both the nature of droplet-technology-derived single-cell expression data and the scale of current and future data sets.

DISCUSSION

Proper removal of batch effects is critical for valid data analysis and interpretation of results. This removal 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, the composition of cell types and states is likely to differ across batches, owing 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 failure can lead to misleading conclusions wherein batch-specific clusters are incorrectly interpreted as distinct cell types. By using both simulated data and real scRNA-seq data sets, we demonstrated that our MNN method is able to successfully remove the batch effect in the presence of differences in composition. Moreover, we demonstrated the MNN method's scalability on large droplet-based data sets.

One prerequisite for our MNN method is that each batch must contain at least one shared cell population with another batch. This requirement is necessary for the correct identification of MNN pairs between batches. Batches without any shared structure are inherently difficult to correct, because the batch effects are completely confounded by biological differences. Such cases provide a motivation for using a 'cell control', 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 batch effects by using a Gaussian kernel. Our method is therefore able to accommodate differences in the size or direction of the batch effect between different cell subpopulations in high-dimensional space. Such differences are not easily handled by methods based on linear models (because they would require explicit modeling of predefined groupings of cells, which would defeat the purpose of using scRNA-seq to study population heterogeneity in the first place). 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 Fig. 7**). An important consequence is that a single control population might not suffice for accurate estimation of local batch effects. Instead, using an appropriately mixed population of cells to properly account for local variation may be necessary.

We 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 that is defined by the first (reference) batch, such that all batches can be analyzed 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 the technical and biological origins of these discrepancies may also be interesting. For instance, one batch might contain cells from a gene-knockout experiment, and the other batch might contain cells from a wild-type organism. In such cases, the correction vectors (provided as an output of the MNN algorithm) could potentially be examined to understand the differences between batches.

Batch correction plays a critical role in the interpretation of scRNA-seq data from both small studies, in which logistical constraints preclude the generation of data in a single batch, and large studies involving international consortia such as the Human Cell Atlas, in which scRNA-seq data are generated for 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 this method will improve the rigor of scRNA-seq data analysis and thus the quality of the biological conclusions.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to F.K. Hamey, J.P. Munro, J. Griffiths and M. Büttner for helpful discussions. L.H. was supported by Wellcome Trust Grant 108437/Z/15 to J.C.M. A.T.L.L. was supported by core funding from CRUK (award number 17197

to J.C.M.). M.D.M. was supported by Wellcome Trust Grant 105045/Z/14/Z to J.C.M. J.C.M. 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 interests.

1. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
2. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
3. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Gierahn, T.M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
5. Hicks, S.C., Townes, F.W., Teng, M. & Irizarry, R.A. Missing data and technical variability in single-cell RNA-sequencing experiments. Preprint at https://www.biorxiv.org/content/early/2017/05/08/025528/ (2017).
6. Tung, P.Y. *et al.* Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
7. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
8. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
9. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
10. Leek, J.T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
11. Spitzer, M.H. *et al.* An interactive reference framework for modeling a dynamic immune system. *Science* **349**, 1259425 (2015).
12. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
13. Scialdone, A. *et al.* Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289–293 (2016).
14. Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.* **3**, 346–360.e4 (2016).
15. Bendall, S.C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
16. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
17. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
18. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
19. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
20. Grün, D. *et al.* De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).
21. Muraro, M.J. *et al.* A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394.e3 (2016).
22. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).
23. Segerstolpe, Å. *et al.* Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
24. Zheng, G.X.Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

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), in which each mixture component represents a different simulated cell type. Two data sets with n = 1,000 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 by 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 hematopoietic data sets. Gene expression counts generated by Nestorowa *et al.*[12] on the SMART-seq2 platform (1,920 cells in total) were downloaded from the NCBI Gene Expression Omnibus (GEO) database, accession number GSE81682. Expression counts generated by Paul *et al.*[18] on the MARS-seq platform (10,368 cells in total) were obtained from NCBI GEO accession GSE72857. Using FACS, the authors identified 2,729 myeloid progenitor cells (CMP, GMP and MEP) as lineage negative (Lin^-) c-Kit^+ $Sca1^-$ and gated the cells further on the basis of the levels of the FcgR and CD34 markers; those cells were used for the analysis in this manuscript. For batch correction, we identified a set of 3,937 highly variable genes in common 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. Cell labels were assigned *a priori* to each cell on the basis of the original publications.

Processing and analysis of the pancreas data sets. Raw data were obtained from NCBI GEO accession numbers GSE81076 (ref. 20) (CEL-seq), GSE85241 (ref. 21) (CEL-seq2) and GSE86473 (ref. 22) (SMART-seq2) and from ArrayExpress accession number E-MTAB-5061 (ref. 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 by using STAR version 2.4.2a[26] with default parameters, and were assigned to Ensembl build 86 protein-coding genes with featureCounts version 1.4.6 (ref. 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, thus leaving a total of 7,236 cells available across all four data sets. Normalization of cell-specific biases was performed for each data set through the deconvolution method of Lun *et al.*[28]. Counts were divided by size factors to obtain normalized expression values that were log transformed after addition of a pseudocount of 1. Highly variable genes were identified in each data set through the method of Brennecke *at al*[25]. We took the union of highly variable genes whose expression was common across all four data sets, thus resulting in 2,507 genes that were used for the MNN batch correction.

Cell-type labels for each data set were assigned on the basis of the provided metadata (GSE86473 and EMTAB-5061) or, if the labels were not provided, were inferred from the data through the method used in the original publication (GSE81076 and 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 neighbor (SNN) graph[29] by using the combined cells and the union of the highly variable genes that were expressed across all data sets. To identify communities of cells, we applied the 'Walktrap' algorithm to the SNN graph[30], with five steps. This procedure identified a total of 11 clusters. To assign specific cell-type labels to those 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 alpha islets, *INS* was used to mark beta islets, *SST* was used to mark delta islets, *PPY* was used to mark gamma islets, *PRSS1* was used to mark acinar cells, *KRT19* was used to mark ductal cells, and *COL1A1* was used to mark mesenchymal 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 by 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, by using the labels derived from the batch-corrected data (described above). We used this design to fit a linear model to the normalized uncorrected log expression values for each gene, then performed empirical Bayes shrinkage to stabilize the sample variances. A moderated t-test was applied to compare the delta and gamma 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. DE 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 with both cell types present. Here, we used a design matrix with a one-way layout constructed from the original cell-type assignments. Delta and gamma 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 10X Genomics droplet-based platform using Chromium v2 chemistry were downloaded from the company website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/). Expression data from 4,459 human T cells (t_4k) and 68,580 PBMCs (pbmc68k) from two separate donors were normalized separately by 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] (PBMCs, 1,409 genes; T cells, 1,219 genes). To define communities of transcriptionally similar cells, we constructed an SNN graph and assigned cells to specific communities by using the Walktrap algorithm. The identity of each community was assigned on the basis of visualization of expression of canonical marker genes in major leukocyte lineages (CD3, CD20, CD14, CD16, CD1C and CD56). Droplet data sets were combined through our MNN approach on the intersection of the two highly variable gene sets (270 genes). Low-dimensional representations of individual and combined data sets were produced with 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., 100% = 68,000 cells. We combined each subset with the set of 4,459 T cells, then recorded the CPU time in the R environment (R Core Team 2017) by using the *system.time* function. For each combination of data, the R environment garbage collector was invoked before the time was recorded.

***t*-SNE plots.** We generated the t-SNE plots by using the Rtsne package with identical parameter settings for the uncorrected data and the data that were batch corrected with MNN, limma and ComBat. In all plots, we used the distance matrix as the input for the Rtsne function (i.e., Rtsne parameter *is_distance* = *TRUE*). For the hematopoietic data, we accounted for the expected continuity of the data structure by choosing a large perplexity parameter (i.e., 90). For all other data sets in which separate clusters were expected to exist, we used the default perplexity parameter (i.e., 30) and again 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 by 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 let $b(i)$ be the average distance of cell i to all cells assigned to the neighboring cluster, i.e., the cluster with the lowest average distance to the cluster of i. The silhouette coefficient for cell i is defined as:

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

doi:10.1038/nbt.4091

A larger $s(i)$ suggests 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. Because dimensionality reduction by t-SNE facilitates more reasonable clustering results than does clustering in high dimensions, we calculated the silhouette coefficients by 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:

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

where x_i is the proportion of cells from batch i in a given region, such that $\Sigma_{i=1}^{c} x_i = 1$. We assessed the total entropy of batch mixing on the first two PCs of the batch-corrected and the uncorrected pancreas data sets, by 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 neighbors for each randomly chosen cell. The total mixing entropy was then calculated as the sum of the regional entropies. We repeated this procedure for 100 iterations with different randomly chosen cells to generate box plots of the total entropy (**Supplementary Figs. 5q** and **6q**).

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 is available on Github (https://github.com/MarioniLab/MNN2017/).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary.

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

25. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
26. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
27. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
28. Lun, A.T., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
29. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).
30. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *ISCIS* **3733**, 284–293 (2005).
31. Buttner, M., Miao, Z., Wolf, A., Teichmann, S.A. & Theis, F.J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. Preprint at https://www.biorxiv.org/content/early/2017/10/09/200345/ (2017).
32. Brandani, G.B. *et al.* Quantifying disorder through conditional entropy: an application to fluid mixing. *PloS One* **6**, e65617 (2013).

nature research

Corresponding author(s): John Marioni

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   All data used in our manuscript are already published data, sample size according to original publications, which is hundreds of cells for each data set.

2. Data exclusions

   Describe any data exclusions.

   We occasionally excluded cells without FACs sorting labels (Paul 2015 data).
   Individual pancreatic cells were excluded if the sequencing depth was low. This is done in the procedure of cell's quality control performed by scran R package for single-cell data processing.
   Genes for which the name could not be matched across the data sets being merged were also excluded from our analysis.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   For demonstrating robustness of our analysis, we performed batch correction on each data set twice, once using all highly variable genes (3904 genes in the haematopoietic data and 2507 in the pancreas data) and once with a set of of 1500 randomly chosen set from highly variable gene set.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   We used all samples in all of the data sets, therefore randomization is not relevant to our study.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   We used all samples in all of the data sets, therefore blinding is not relevant to our study.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

Nature Biotechnology: doi:10.1038/nbt.4091

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	R codes for the MNN batch correction method is available in package "scran" (version 1.6.6). All codes used for the analysis and generating of the results (including the call to required packages STAR, Rtsne, destiny, scran, kBET, limma, ComBat, etc) are deposited at: https://github.com/MarioniLab/MNN2017

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	We used published data. GEO accession numbers: GSE81682 and GSE72857 for haematopoietic ata sets. GSE81076, GSE85241, GSE86473 and the ArrayExpress accession number E-MTAB-5061 for the pancreas data sets. 10X (droplet) data is publicly available and were downloaded directly from the 10X website.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used in this study.
b. Describe the method of cell line authentication used.	-
c. Report whether the cell lines were tested for mycoplasma contamination.	-
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	-

Nature Biotechnology: doi:10.1038/nbt.4091

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	-No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	-No human research participants were used in this study.

Nature Biotechnology: doi:10.1038/nbt.4091