Fast, sensitive, and flexible integration of single cell data with Harmony

Ilya Korsunsky[1234], Jean Fan[5], Kamil Slowikowski[123], Fan Zhang[1234], Kevin Wei[2], Yuriy Baglaenko[1234], Michael Brenner[2], Po-Ru Loh[134], Soumya Raychaudhuri[12346]

[1]Center for Data Sciences, Brigham and Women's Hospital, Massachusetts, USA. [2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, [3]Department of Biomedical Informatics, Harvard Medical School, Massachusetts, USA. [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, USA. [6] Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK.

* Correspondence to:

Soumya Raychaudhuri
77 Avenue Louis Pasteur, Harvard New Research Building, Suite 250D
Boston, MA 02446, USA.
soumya@broadinstitute.org; 617-525-4484 (tel); 617-525-4488 (fax)

Abstract

The rapidly emerging diversity of single cell RNAseq datasets allows us to characterize the transcriptional behavior of cell types across a wide variety of biological and clinical conditions. With this comprehensive breadth comes a major analytical challenge. The same cell type across tissues, from different donors, or in different disease states, may appear to express different genes. A joint analysis of multiple datasets requires the integration of cells across diverse conditions. This is particularly challenging when datasets are assayed with different technologies in which real biological differences are interspersed with technical differences. We present Harmony, an algorithm that projects cells into a shared embedding in which cells group by cell type rather than dataset-specific conditions. Unlike available single-cell integration methods, Harmony can simultaneously account for multiple experimental and biological factors. We develop objective metrics to evaluate the quality of data integration. In four separate analyses, we demonstrate the superior performance of Harmony to four single-cell-specific integration algorithms. Moreover, we show that Harmony requires dramatically fewer computational resources. It is the only available algorithm that makes the integration of ~10^6 cells feasible on a personal computer. We demonstrate that Harmony identifies both broad populations and fine-grained subpopulations of PBMCs from datasets with large experimental differences. In a meta-analysis of 14,746 cells from 5 studies of human pancreatic islet cells, Harmony accounts for variation among technologies and donors to successfully align several rare subpopulations. In the resulting integrated embedding, we identify a previously unidentified population of potentially dysfunctional alpha islet cells, enriched for genes active in the Endoplasmic Reticulum (ER) stress response. The abundance of these alpha cells correlates across donors with the proportion of dysfunctional beta cells also enriched in ER stress response genes. Harmony is a fast and flexible general purpose integration algorithm that enables the identification of shared fine-grained subpopulations across a variety of experimental and biological conditions.

Introduction

Recent technological advances[1] have enabled unbiased single cell transcriptional profiling of thousands of cells in a single experiment. Projects such as the Human Cell Atlas[2] (HCA) and Accelerating Medicines Partnership[3,4] exemplify the growing body of reference datasets of primary human tissues. While individual

experiments contribute incrementally to our understanding of cell types, a comprehensive catalogue of healthy and diseased cells will require the integration of multiple datasets across donors, studies, and technological platforms. Moreover, in translational research, joint analyses across tissues and clinical conditions will be essential to identify disease expanded populations. Without effective strategies, single cell RNA-seq from different studies appear to be hopelessly confounded by data source[5]. Recognizing this key issue, investigators have developed unsupervised multi-dataset integration algorithms, such as Seurat MultiCCA[6], MNN Correct[7], Scanorama[8], and BBKNN[9] to enable joint analysis. These methods embed cells from diverse experimental conditions and biological contexts into a common reduced dimensional embedding to enable shared cell type identification.

We introduce Harmony for multi-dataset integration, to meet three key challenges of unsupervised scRNAseq joint embedding. First, cell types with regulatory or pathogenic roles are often rare, with subtle transcriptomic signatures. Integration must be able to identify such rare cell types, particularly those whose subtle signatures are initially obscured by other technical or biological confounders. To be sensitive to subpopulations with subtle signatures, Harmony uses a two-step iterative strategy that removes the effect of such confounding factors at each round. This makes it easier to identify shared cell types whose expression signatures were obscured in the original data. Second, the number of cells in experiments is quickly expanding, exceeding 100,000 cells in atlas-like datasets. Integration algorithms must scale computationally, both in terms of runtime and required memory resources. To scale to big data, Harmony uses linear methods and avoids costly cell-to-cell comparisons. Third, more complex experimental design of single cell analysis compares cells from different donors, tissues, and technological platforms. In order for the joint embedding to be free of the influence of each, integration must simultaneously account for multiple sources of variation. The Harmony clustering objective function is formulated to account for any number of discrete covariates. Harmony is available as an R package on github (https://github.com/immunogenomics/harmony), with functions for standalone and Seurat[6] pipeline analyses.

Here, we demonstrate how Harmony address the three unmet needs outlined above. First, we give an overview of the Harmony algorithm. Then, we then integrate cell line datasets and introduce a metric to quantify the cell-type accuracy and degree of dataset-mixing before and after integration. All measures of cell-type accuracy are based on annotation within datasets separately. As Harmony does not use cell-type

information to integrate cells, these labels provide an unbiased quantification of accuracy. We then demonstrate that Harmony scales more effectively than any of the other available algorithms across a range of dataset sizes, from 25,000 to 500,000 total cells. We then use Harmony to identify shared cell in three PBMC datasets with large technical differences. Finally, in a pancreatic islet cell meta-analysis, we demonstrate the power of Harmony to simultaneously integrate donor- and technology-specific effects by identifying a putative novel pancreatic islet cell subtype.

Results

Harmony Iteratively Learns a Cell-Specific Linear Correction Function

Starting with a PCA embedding, Harmony first groups cells into multi-dataset clusters (**Figure 1A**). We use soft clustering, assigning cells to potentially multiple clusters, to account for smooth transitions between cell states. In our thinking, these clusters serve as surrogate variables, rather than actually defining discrete cell-types. We developed a novel variant of soft k-means clustering to favor clusters with representation across multiple datasets (**Online Methods**). Clusters containing cells that are disproportionately represented by a small subset of datasets are penalized by an information theoretic metric. Harmony can employ multiple cluster penalties if there are multiple technical or biological factors, such as different batches and different technology platforms. Soft clustering preserves discrete and continuous topologies while avoiding local minima that might result from too quickly maximizing representation across multiple datasets, and preserves uncertainty. After clustering, each dataset has a cluster-specific centroid (**Figure 1B**) that is used to compute cluster-specific linear correction terms (**Figure 1C**). Under favorable conditions, the surrogate variables, defined by cluster membership, correspond to cell types and cell states. Thus, the cluster-specific correction factors that Harmony computes correspond to individual cell-type and cell-state specific correction factors. In this way, Harmony learns a simple linear adjustment function that is sensitive to intrinsic cellular phenotypes. Finally, each cell is assigned a cluster-weighted average of these terms and corrected by its cell-specific linear factor (**Figure 1D**). As a result, each cell has a potentially unique correction factor, depending on its soft clustering distribution. Harmony iterates these four steps until convergence. At convergence, additional iterations would assign cells to the same clusters and compute the same linear correction factors.

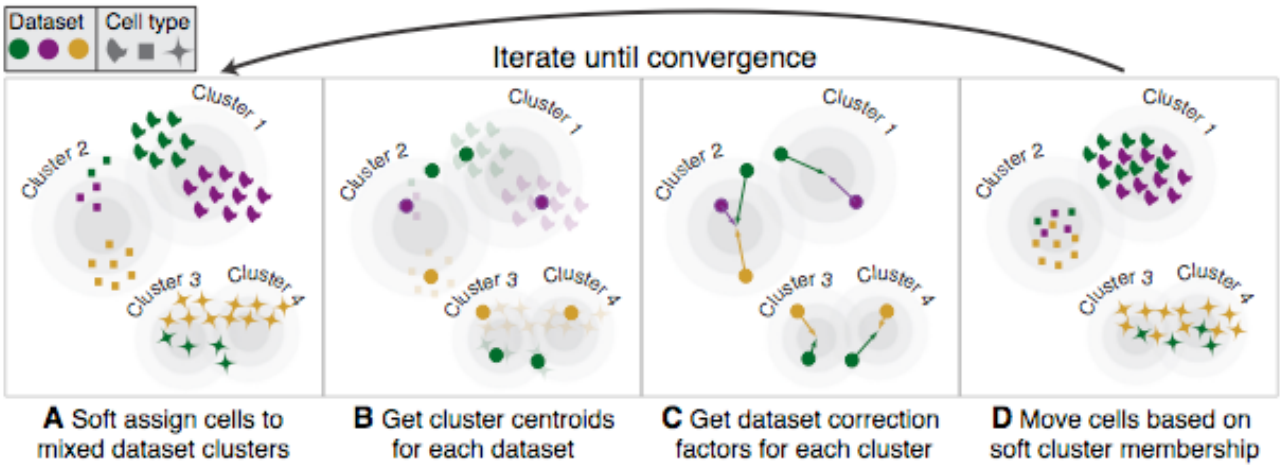


Figure 1. Overview of Harmony algorithm. We represent datasets represented by colors, and different cell types by shapes. First, Harmony applies principal components analysis to embed transcriptome wide expression profiles into reduced dimensional space. Then we apply an iterative process to remove the data set specific effects. (**A**) Harmony assigns cells probabilistically to clusters, maximizing the diversity of datasets within each cluster. (**B**) Harmony calculates a global centroid across all datasets for each cluster, as well as dataset-specific centroids. (**C**) Within each cluster, Harmony calculates a correction factor for each dataset based on the centroids. (**D**) Finally, Harmony corrects each cell with a cell-specific factor based on C. Since Harmony uses soft clustering, each individual cell may be corrected by a linear combination of multiple factors proportion to its soft cluster assignments made in A. Harmony repeats steps A to D until convergence. The dependence between cluster assignment and dataset diminishes with each round.

Quantification of Error and Accuracy in Cell Line Integration

We first assessed Harmony using well-annotated datasets, in order to evaluate performance on both integration (mixing of datasets) and accuracy (no mixing of cell types). Both metrics are important. Perfect integration can be achieved by simply mixing all cells, regardless of cellular identity. Similarly, high accuracy can be achieved by partitioning cell types into broad clusters without mixing datasets in small neighborhoods. In this situation, broad cellular states are defined, but fine-grained cellular substates and subtypes are confounded by the originating dataset. In order to quantify integration and accuracy of this embedding we felt that it was important that we have an objective metric. To this end, we compute the Local Inverse Simpson's Index (LISI, **Online Methods**) in the local neighborhood of each cell. To assess integration, we employ *integration LISI* (iLISI, **Figure 2A**), denotes the effective number of datasets in a neighborhood. Neighborhoods represented by only a single dataset get an iLISI of 1, while neighborhoods with an equal number of cells from 2 datasets get an iLISI of 2. Note that even under ideal mixing, if the datasets have

22 different numbers of cells, iLISI would be less than 2. To assess accuracy, we use *cell-type LISI* (cLISI,
23 **Figure 2B**), the same mathematical measure, but applied to cell-type instead of dataset labels. Accurate
24 integration should maintain a cLISI of 1, reflecting a separation of unique cell types throughout the
25 embedding. An erroneous embedding would include neighborhoods with a cLISI of 2, reflecting the grouping
26 of different cell types.

27 We begin with three datasets from two cells lines: (1) pure Jurkat, (2) pure 293T and (3) a 50:50
28 mix[10]. These datasets are particularly ideal for illustration and for assessment, as each cell can be
29 unambiguously labeled Jurkat or 293T (**Figure S1**). A thorough integration would mix the 1799 Jurkat cells
30 from the mixture dataset with 3255 cells from the pure Jurkat dataset and the 1565 293T cells from the
31 mixture dataset with the 2859 from the pure 293T dataset. Thus, we expect the average iLISI to range from 1,
32 reflecting no integration, to 1.8 for Jurkat cells and 1.5 for 293T cells[1], reflecting maximal accurate integration.
33 Application of a standard PCA pipeline followed by UMAP embedding demonstrates that the cells group
34 broadly by dataset and cell type. This is both visually apparent (**Figure 2C,D**) and quantified (**Figure 2E,F**)
35 with high accuracy reflected by a low cLISI (median iLISI 1.00, 95% [1.00, 1.00]). However the iLISI (median
36 iLISI 1.01, 95% [1.00, 1.61]) is also low, reflecting imperfect integration, and ample structure within each cell-
37 type reflecting the data set of origin. After Harmony, cells from the 50:50 dataset are mixed into the pure
38 datasets (**Figure 2E**), which is appropriate in this case since these cell-lines have no additional biological
39 structure. The increased iLISI (**Figure 2D,** median iLISI 1.59, 95% [1.27, 1.97]) reflects the mixing of
40 datasets, while the low cLISI (median iLISI 1.00, 95% [1.00, 1.02]) reflects the accurate separation of Jurkat
41 from 293T cells. iLISI and cLISI provide a quantitative way to assess the integration and accuracy of multiple
42 algorithms. We repeated the integration and LISI analyses with MNN Correct, BBKNN, MultiCCA, and
43 Scanorama (**Figure 2G**). While Harmony had the highest iLISI, Scanorama MNN Correct, and BBKNN
44 provided lower levels of integration, evidence by lower iLISI. MultiCCA actually separated previously mixed
45 datasets, yielding a lower iLISI (median 1.00, 95% [1.00, 1.28]) than before integration. Except for MultiCCA
46 (median cLISI 1.00, 95% [1.00, 1.23]), the other algorithms maintained high accuracy (**Table S1,** median
47 cLISI 1.00, 95% [1.00, 1.00]).

---

[1] $1.5 = 1 / [(1565 / (1565 + 2859))^2 + (2859/ (1565 + 2859))^2]$, $1.8 = 1 / [(1799 / (1799 + 2859))^2 + (3255 / (1799 + 3255))^2]$

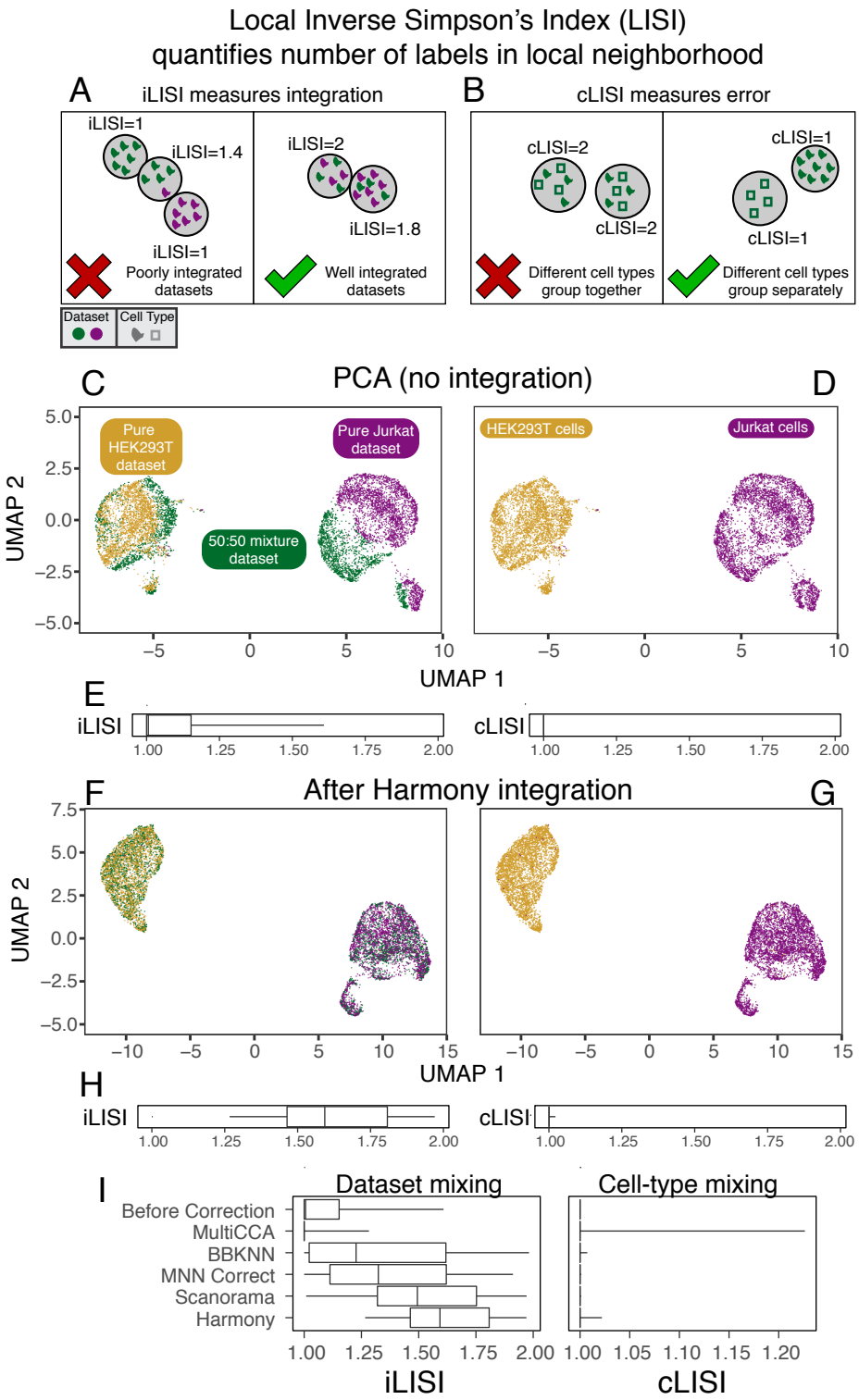


Figure 2. Quantitative assessment of dataset mixing and cell-type accuracy with cell line datasets. (**A**) iLISI measures the degree of mixing among datasets in an embedding, ranging from 1 in an unmixed space to B in a well mixed space. B is the number of datasets in the analysis. (**B**) cLISI measures integration accuracy using the same formulation but computed on cell-type labels instead. An accurate embedding has a cLISI close to 1 for every neighborhood, reflecting separation of different cell types. Jurkat and HEK293T cells from pure (purple and yellow) and mixed (green) cell-line datasets were analyzed together. Before Harmony integration, cells

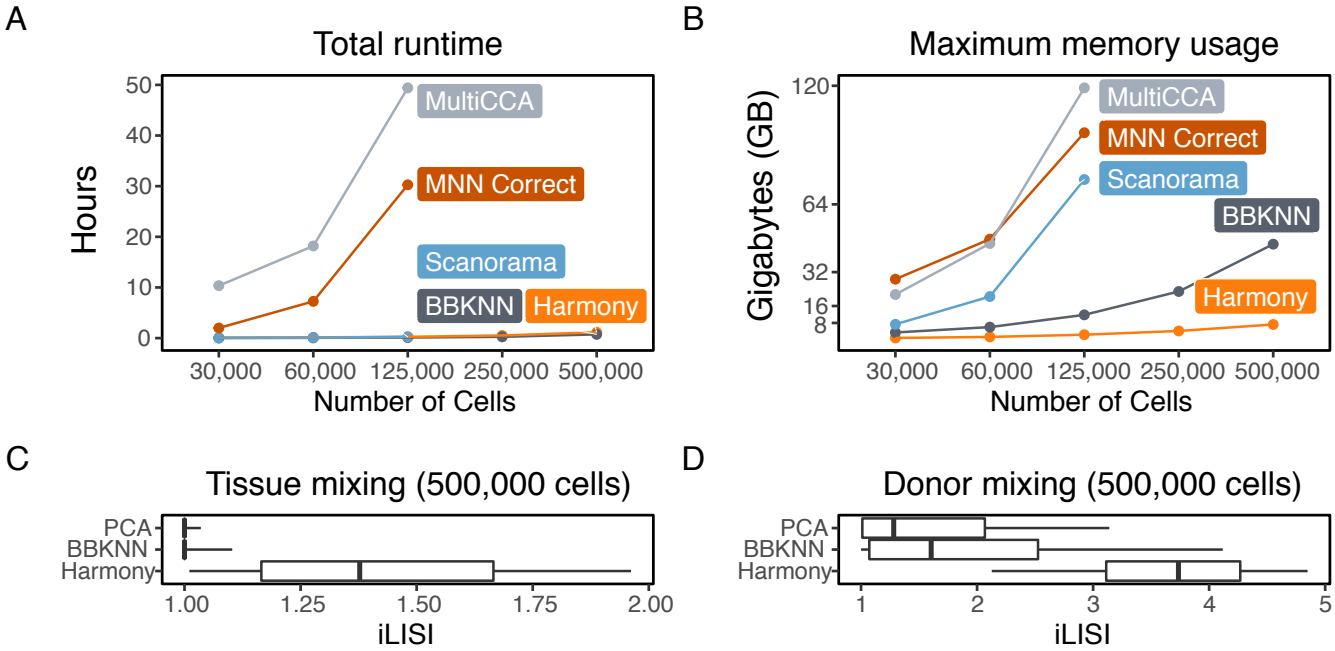
54 grouped by dataset (**C**) and known cell-type (**D**). iLISI and cLISI (**E**) were computed for every cell's neighborhood and summarized
55 with quantiles (5, 25, 50, 75, 95). After Harmony integration, cells from the mixture dataset are mixed into the other datasets (**F**),
56 achieved by mixing Jurkat with Jurkat cells and HEK293T with HEK293T cells (**G**). iLISI and cLISI (**H**) were re-computed in the
57 Harmony embedding. (**I**) This analysis was repeated for other algorithms and compared against no integration and Harmony
58 integration using iLISI and cLISI quantiles.

59 Harmony Scales to Enable Analysis of Large Data

60 As an integral part of the scRNAseq analysis pipeline, the integration algorithm must run in a reasonable
61 amount of time and within the memory constraints of standard computers. To this end, we evaluated the
62 computational performance of Harmony, measuring both the total runtime and maximal memory usage. To
63 demonstrate the scalability of Harmony versus other methods, we downsampled 528,688 cells from 16
64 donors in the HCA data[11] to create 5 benchmark datasets with 500,000, 250,000, 125,000, 60,000, and
65 30,000 cells. We reported the runtime (**Table S2**) and memory (**Table S3**) for all benchmarks. Harmony
66 runtime scaled well for all datasets (**Figure 3A**), ranging from 4 minutes on 30,000 cells to 68 minutes on
67 500,000 cells. It was 30 to 200 times faster than MultiCCA and MNN Correct. The runtimes for Harmony,
68 BBKNN, and Scanorama were comparable for datasets with up to 125,000 cells. Harmony required very little
69 memory (**Figure 3B**) compared to other algorithms, only 0.9GB on 30,000 cells and 7.2GB on 500,000 cells.
70 At 125,000 cells, Harmony required 30 to 50 times less memory than Scanorama, MNN Correct and Seurat
71 MultiCCA; these other methods could not scale to 500,000 cells. Notably, BBKNN was the only other
72 algorithm able to finish on the 500,000 cell dataset, taking 44 minutes and 45GB of RAM. However, the
73 BBKNN embedding barely integrated tissues (**Figure 3C**, median iLISI 1.00, 95% [1.00, 1.10]) or donors
74 (**Figure 3D**, median iLISI 1.60, 95% [1.00, 4.11]) above PCA alone (tissue median iLISI 1.00, 95% [1.00,
75 1.03], donor median iLISI 1.38, 95% [1.00, 3.14]).

76 Importantly, in addition to better computational performance of other algorithms Harmony returned a
77 substantially more integrated space that other competing algorithms, allowing for the identification of shared
78 cell types across tissues (**Table S4,** median iLISI 1.40, 95% [1.04, 1.97] compared to medians of 1.00 to
79 1.12) and donors (median iLISI 3.93, 95% [2.46, 4.95] compared to medians of 1.07 to 2.82). This evidence
80 of high computational efficiency and effective integration, suggests that Harmony could analyze very large

81 datasets (10^5 – 10^6 cells) on personal computers, without the need for specialized machines. Alternative
82 methods may require extensive parallelization to run modestly sized datasets.



83

84 **Figure 3.** Computational efficiency benchmarks. We ran Harmony, BBKNN, Scanorama, MNN Correct, and MultiCCA on 5
85 downsampled HCA datasets of increasing sizes, from 25,000 to 500,000 cells. We recorded the (**A**) total runtime and (**B**) maximum
86 memory required to analyze each dataset. Scanorama, MultiCCA, and MNN Correct were terminated for excessive memory requests
87 on the 250,000 and 500,000 cell datasets. For Scanorama and Harmony, we quantified the extent of integration in the 500,000 cell
88 benchmark for (**C**) the two tissues (cord blood and bone marrow) and (**D**) the 16 donors. For reference, (**C**) and (**D**) report the iLISI
89 scores prior to integration.

90 Deep Integration Enables Identification of Broad and Fine-Grained PBMCs
91 Subpopulations

92 To assess how effective Harmony might be under more challenging scenarios, we gathered three datasets of
93 human PBMCs, each assayed on the Chromium 10X platform but prepared with different protocols: 3-prime
94 end v1 (3pV1), 3-prime end v2 (3pV2), and 5-prime (5p) end chemistries. Pooling all the cells together, we
95 performed a joint analysis. Before integration, cells group primarily by dataset (**Figure 4A**, median iLISI 1.00,
96 95% [1.00, 1.00]). Harmony integrates the three datasets (**Figure 4B,** median iLISI 1.96, 95% [1.36, 2.56]),
97 considerably more than do other methods (**Figure 4C**). To assess accuracy, within each dataset, we

separately annotated (**online methods**) broad cell clusters with canonical markers of major expected populations (**Figure S2**): monocytes (CD14+ or CD16+), dendritic cells (FCER1A+), B cells (CD20+), T cells (CD3+), Megakaryocytes (PPBP+), and NK cells (CD3-/GNLY+) before clustering. We observed that Harmony retained differences among cell types (**Figure 4D** median cLISI 1.00, 95% [1.00, 1.02]). The greater dataset integration, compared to other algorithms, affords a unique opportunity to identify fine-grained cell subtypes. Using canonical markers (**Figure 4E**), we identified shared subpopulations of cells (**Figure 4F**) including naive CD4 T (CD4+/CCR7+), effector memory CD4 T (CD4+/CCR7-), [SR2] Treg (CD4+/FOXP3+), memory CD8 (CD8+/GZMK-), effector CD8 T (CD8+/GZMK+), naive B (CD20+/CD27-), and memory B cells (CD20+/CD27+). In the embeddings produced by other algorithms, the median iLISI did not exceed 1.1 (**Table S5**). Accordingly, the subtypes identified with Harmony reside in dataset-specific, rather than dataset-mixed clusters (**Figure S3**). In practice, Harmony provides a uniquely fine-grained embedding for the unbiased discovery of both broad cell types and fine-grained subpopulations.

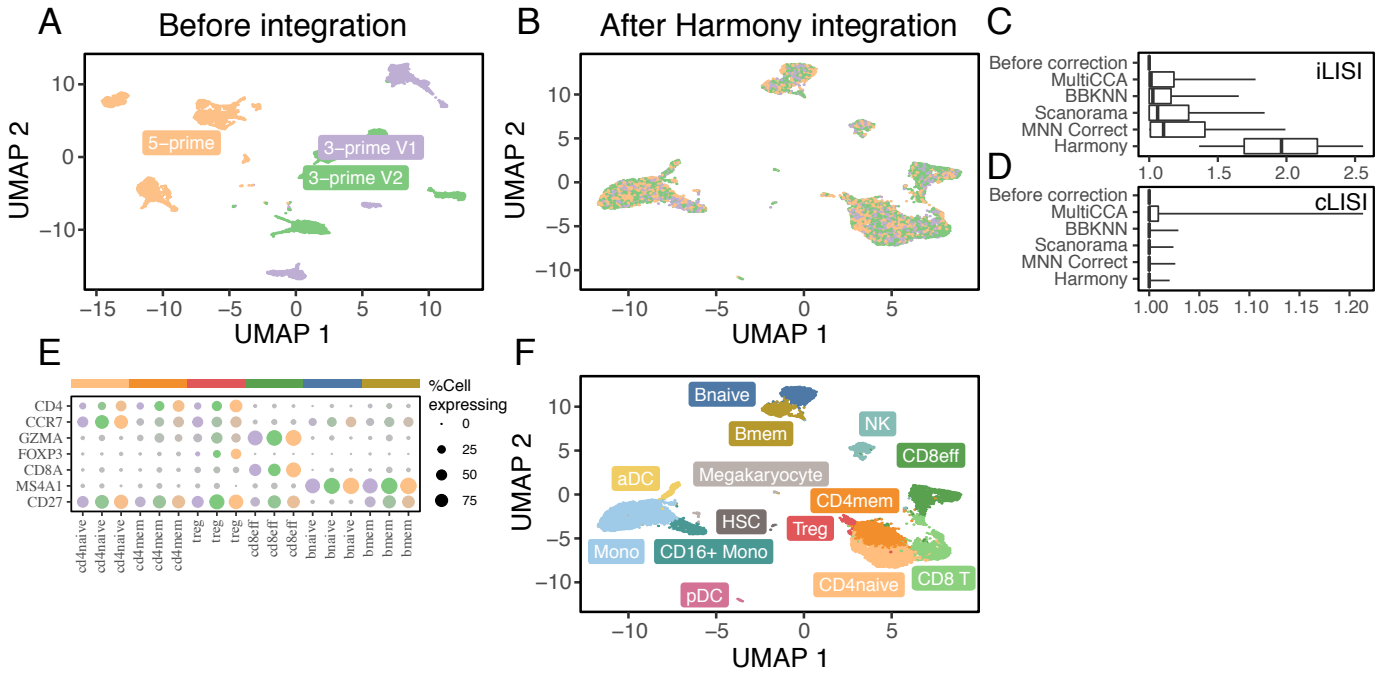


Figure 4. Fine-grained subpopulation identification in PBMCs across technologies. Three PBMC datasets were assayed with 10X, using different library construction protocols: 5-prime (orange), 3-prime V1 (purple), and 3-prime V2 (green). Before integration (**A**), cells group by dataset. After Harmony integration (**B**), datasets are mixed together. (**C**) Harmony achieves the most thorough integration among datasets, while preserving (**D**) cell type differences. Using canonical markers (**E**), we identified (**F**) 5 shared subtypes of T cells and 2 shared subtypes of B cells. (**G**) Other integration algorithms fail to group these cells by subtype.

Simultaneous Integration Across Donors and Technologies Identifies Rare Pancreas Islet Subtypes

We considered a more complex experimental design, in which integration must be performed simultaneously over more than one variable. We gathered human pancreatic islet cells from independent five studies[12–16], each of which were generated with a different technological platform. Integration across platforms is already challenging. However, within two of the datasets[12,14], the authors also reported significant donor-specific effects. In this scenario, a successful integration of these studies must account for the effects of both technologies and donors, which may both affect different cell types in different ways. Harmony is the only single cell integration algorithm that is able to integrate over more than one variable, hence we omit a comparison against other methods.

As before, we assess cell type accuracy cLISI with canonical cell types identified separately within each dataset (**Figure S4**): alpha (GCG+), beta (MAFA+), gamma (PPY+), delta (SST+), acinar (PRSS1+), ductal (KRT19+), endothelial (CDH5+), stellate (COL1A2+), and immune (PTPRC+). Since there are two integration variables, we asses both donor iLISI and technology iLISI. Prior to integration, cells are separated by technology (**Figure 5A,E**, median iLISI 1.00 95% [1.00, 1.06]), donor (**Figure 5B,E**, median iLISI 1.42 95% [1.00, 5.50]), and cell type (**Figure 5E**, median cLISI 1.00 95% [1.00, 1.48]). The wide range of donor-iLISI reflects that in the CEL-seq, CEL-seq2, and Fluidigm C1 datasets, many donors were well mixed prior to integration. Harmony integration mixes cells by both technology (**Figure 5C,E**, median iLISI 2.27 95% [1.31, 3.27]) and donor (**Figure 5D,E**, median iLISI 4.71 95% [1.81, 6.36]).

Harmony was able to discern rare cell subtypes (**Figure 5F**) across the 5 datasets (**Figure 5G**). We labeled previously described subtypes using canonical markers: activated stellate cells (PDGFRA+), quiescent stellate cells (RGS5+), mast cells (BTK+), macrophages (C1QC+), and beta cells under endoplasmic reticulum (ER) stress (**Figure 5H**). Beta ER stress cells may represent a dysfunctional population. This cluster has significantly lower expression of genes key to beta cell identify[17] and function[18]: PDX1, MAFA, INSM1, NEUROD1 (**Figure 5I**). Further, Sachdeva et al[19] suggest that PDX1 deficiency makes beta cells less functional and exposes them to ER stress induced apoptosis.

Intriguingly, we also observed an alpha cell subset that to our knowledge was not previously described. This cluster was also enriched with genes involved in ER stress (**Figure 5J**, DDIT3, ATF3, ATF4,

and HSPA5). Similar to the beta ER stress population, these alpha cells also expressed significantly lower levels of genes necessary for proper function[20,21]: GCG, ISL1, ARX, and MAFB (**Figure 5K**). A recent study reported ER stress in alpha cells in mice and linked the stress to dysfunctional glucagon secretion[22]. Moreover, we found that the proportions of alpha and beta ER stress cells are significantly correlated (spearman r=0.46, p=0.004, **Figure 5L**) across donors in all datasets. These results suggest a basis for alpha cell injury that might parallel beta cell dysfunction in humans during diabetes[23].

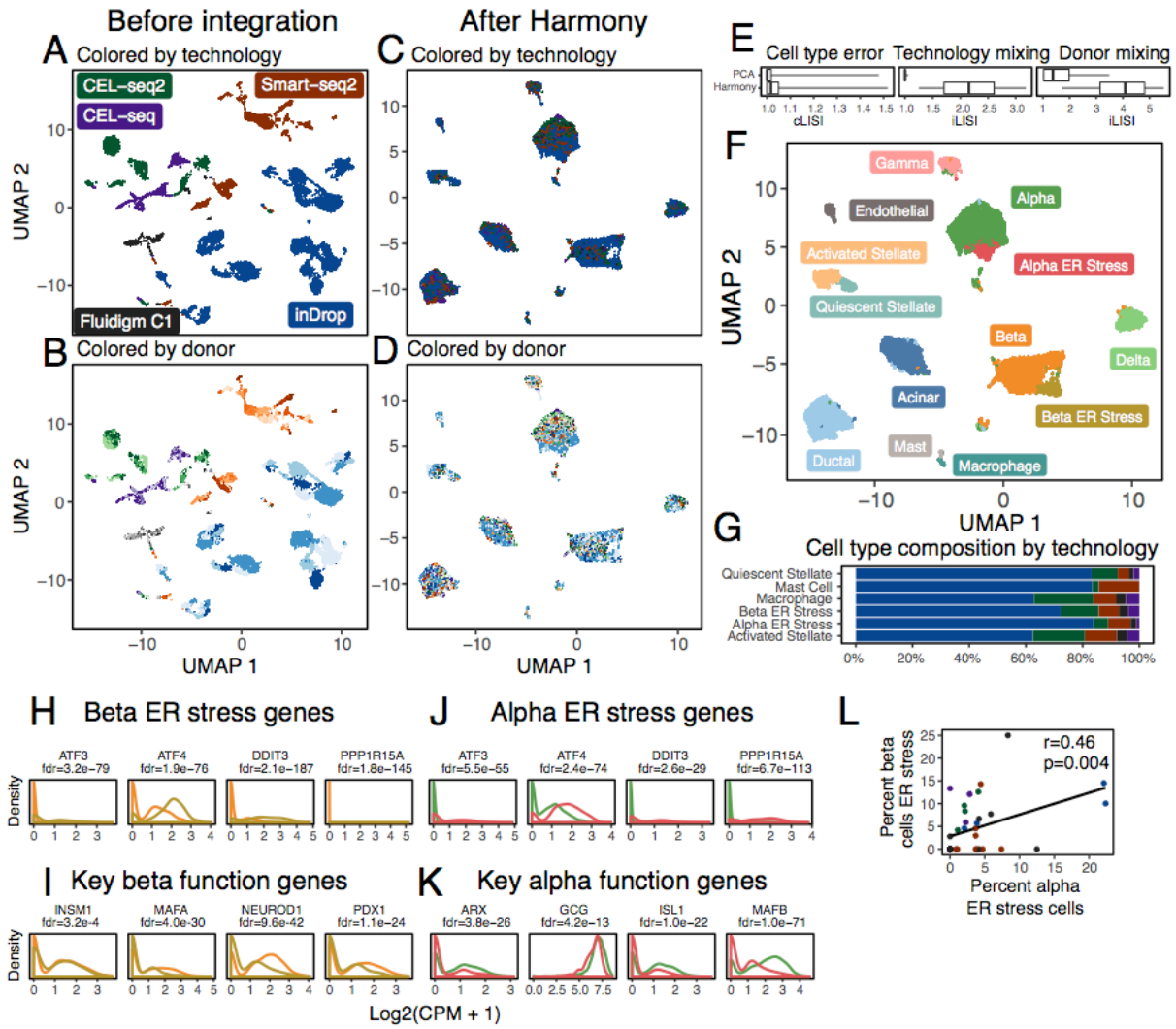


Figure 5. Integration of pancreatic islet cells by both donor and technology. Human pancreatic islet cells from 36 donors were assayed on 5 different technologies. Cells initially group by (**A**) technology, denoted by different colors, and (**B**) donor, denoted by shades of colors. Harmony integrates cells simultaneously across (**C**) technology and (**D**) donor. Integration across both variables was quantified with iLISI (**E**) and error was computed with cLISI (**E**). Clustering in the Harmony embedding identified common and rare cell types, including a previously undescribed alpha population. Except for activated stellate cells, all rare cell types were found across the 5 technology datasets (**G**). The new alpha cluster was enriched for ER stress genes (**I**), just like the previously identified beta ER stress cluster (**J**). The abundances of the two ER stress populations were correlated across donors (**H**). Key genes

58 necessary for endocrine function were downregulated in the alpha (**K**) and beta (**L**) ES stress clusters.

59

60 Discussion

61 We showed that Harmony address the three key challenges we laid out for single cell integration analyses:
62 scaling to large datasets, identification of both broad populations and fine-grained subpopulations, and
63 flexibility to accommodate complex experimental design. We evaluated the degree of mixing among datasets
64 using a quantitative metric, the iLISI. Apart from its use benchmarking, iLISI was particularly important in
65 analyses with more than 3 datasets. Here, we observed that the commonly utilized approach of assessing
66 integration visually was subjective and insensitive, particularly when the number of samples, batches or cell
67 types was large. iLISI provided a quantitative and interpretable metric to help guide analysis. In the
68 computational efficiency benchmarks, we found that 3 of the 5 algorithms were not able to scale beyond
69 125,000 cells because they exceeded the memory resources of our 128GB servers. We were struck by the
70 fact many researchers routinely analyze data on personal computers, which often do not exceed 8 or 16GB.
71 Harmony, which only required 7.2GB to integrate 500,000 cells, is the only algorithm that would enable the
72 integration of large datasets on personal computers. With the pancreatic islet meta-analysis, we
73 demonstrated that Harmony is able to account simultaneously for donor and technology specific effects. One
74 solution to this multi-level problem stepwise is to first globally regress out one variable from the gene
75 expression values and then performing single cell integration on the resulting expression matrix. Harmony
76 allows for cell-type aware integration of both variables, simultaneously avoiding global correction terms that
77 treat all cell uniformly. However, the global regression strategy is flexible enough to account for continuous
78 variables, such as read depth or cell quality. In the future, Harmony should also be able to account for such
79 non-discrete sources of variation.

80 We noticed that it is not an uncommon practice to apply a batch-sensitive gene scaling step step
81 before using a single-cell integration algorithm. Specifically, many investigators scale gene expression values
82 within datasets separately, before pooling cells into a single matrix. We show (**Supplementary Results**) that
83 this strategy may make it easier to integrate datasets (**Figure S5A,B**) in the rare situation in which the all cell
84 populations and subpopulations are present across all analyzed datasets. However, when the datasets

consist of overlapping but not identical populations, this scaling strategy is less effective (**Figure S5C**) and may indeed even increase error (**Figure S5D**). For this reason, we do not use this scaling strategy in this manuscript. As part of a universal pipeline, Harmony finds highly integrated embeddings without the need for within-dataset scaling.

Harmony allows the user to define a hyperparameter for each covariate that guides how aggressively to integrate out each source of variation. When the penalty is 0, Harmony performs minimal integration. Curiously, we noticed that for small inter-datasets differences, cells from multiple datasets cluster together without the penalty. In this case, Harmony still integrates the cells during the linear correction phase. On the other hand, one could imagine that with an infinitely large penalty hyperparameter, Harmony would overmix datasets during clustering and hence overcorrect the data. We evaluated the effect of the diversity penalty in the PBMCs example (see **Supplementary Results**), and observed that the Harmony embedding is robust to a wide range of penalties (**Figure S6**). Nonetheless, as with any integration algorithm, we urge the user to understand the effects of hyperparameters and experiment with several values.

Harmony is designed to input PCA cell coordinates and covariate labels as input and then output integrated cell coordinates. As such, Harmony should be used as an upstream step in a full analysis pipeline. Downstream analyses, such as clustering, trajectory analysis, and visualization, can use the integrated Harmony embeddings as they usually would PCA coordinates. As a corollary, Harmony does not currently alter the expression values of individual genes to account for dataset-specific differences. We recommend using a batch-aware approach, such as a linear model with covariates, for differential expression analysis.

In our meta-analysis of pancreatic islet cells, we identified a previously undescribed rare subpopulation of alpha ER stress cells **(Figure 5F,J)**. Similar to beta ER stress cells, they appear to have reduced endocrine function (**Figure 5K**). Because Harmony integrated over both donors and technology (**Figure 5C,D,E**), we were able to identify the significant association between the proportion of alpha to beta ER stress populations across donors (**Figure L**). Based on this correlation and similar stress response patterns, it is possible that these two populations are involved in a coordinated response to an environmental stress. Beta cell dysfunction is key to the pathogenesis of diabetes[23]. Experimental follow up on this alpha subtype and its relation to beta ER stress cells may yield insight into disease. This analysis demonstrates the

power of Harmony's multilevel integration to mix diverse datasets and uncover potentially novel rare cell types.

Acknowledgements

This work was supported in part by funding from the National Institutes of Health (UH2AR067677, U19AI111224, and 1R01AR063759 (to S.R.)) and T32 AR007530-31. We thank members of the Raychaudhuri and Brenner labs for comments and discussion.

19 # References

20 1. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the
21 past decade. *Nat. Protoc.* **13,** 599–604 (2018).
22 2. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6,** (2017).
23 3. Zhang, F., Wei, K., Slowikowski, K., Fonseka, C. Y. & Rao, D. A. Defining Inflammatory Cell States in
24 Rheumatoid Arthritis Joint Synovial Tissues by Integrating Single-cell Transcriptomics and Mass
25 Cytometry. *bioRxiv* (2018). doi:10.1101/351130
26 4. Arazi, A. *et al.* The immune cell landscape in kidneys of lupus nephritis patients. *bioRxiv* 363051 (2018).
27 doi:10.1101/363051
28 5. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell
29 RNA-sequencing experiments. *Biostatistics* **19,** 562–578 (2017).
30 6. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data
31 across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).
32 7. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing
33 data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).
34 8. Hie, B. L., Bryson, B. & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data.
35 *bioRxiv* 371179 (2018). doi:10.1101/371179
36 9. Park, J.-E., Polanski, K., Meyer, K. & Teichmann, S. A. Fast Batch Alignment of Single Cell
37 Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. *bioRxiv* 397042
38 (2018). doi:10.1101/397042
39 10. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8,**
40 14049 (2017).
41 11. Li, B. *et al.* HCA Data Portal - Census of Immune Cells. doi:https://preview.data.humancellatlas.org.
42 12. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and
43 Intra-cell Population Structure. *Cell Syst.* **3,** 346–360.e4 (2016).
44 13. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-
45 specific expression changes in type 2 diabetes. *Genome Res.* **27,** 208–222 (2017).

46 14. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type
47 2 Diabetes. *Cell Metab.* **24,** 593–607 (2016).
48 15. Grün, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem*
49 *Cell* **19,** 266–277 (2016).
50 16. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3,** 385–394.e3
51 (2016).
52 17. Gao, T. *et al.* Pdx1 maintains β cell identity and function by repressing an α cell program. *Cell Metab.* **19,**
53 259–271 (2014).
54 18. Jia, S. *et al.* Insm1 cooperates with Neurod1 and Foxa2 to maintain mature pancreatic β-cell function.
55 *EMBO J.* **34,** 1417–1433 (2015).
56 19. Sachdeva, M. M. *et al.* Pdx1 (MODY4) regulates pancreatic beta cell susceptibility to ER stress. *Proc.*
57 *Natl. Acad. Sci. U. S. A.* **106,** 19090–19095 (2009).
58 20. Katoh, M. C. *et al.* MafB Is Critical for Glucagon Production and Secretion in Mouse Pancreatic α Cells In
59 Vivo. *Mol. Cell. Biol.* **38,** (2018).
60 21. Liu, J. *et al.* Islet-1 Regulates Arx Transcription during Pancreatic Islet α-Cell Development. *J. Biol.*
61 *Chem.* **286,** 15352–15360 (2011).
62 22. Akiyama, M. *et al.* X-box binding protein 1 is essential for insulin regulation of pancreatic α-cell function.
63 *Diabetes* **62,** 2439–2449 (2013).
64 23. Burcelin, R., Knauf, C. & Cani, P. D. Pancreatic alpha-cell dysfunction in diabetes. *Diabetes Metab.* **34**
65 **Suppl 2,** S49–55 (2008).

66

67 # Figure Legends

68
69
70
71
72