Supplementary Table S1. Methods for the integration of scRNA-seq datasets

	Dimension reduction[a]	Similarity search	Batch correction	Output type (G/E/W)[d]	Speed	Peak memory use	GPU support	Bio-conservation vs Batch correction[1]	Performance	Reference
limma	-	-	LD	G	High[2]					[3]
ComBat	-	-	LD	G	High[1,2,4]				High[5]	[6]
ZINB-WaVE	-[b]	-	LD	E	Low[2]	High[2]				[7]
MNN	-	MNN	MNN–GK	G	Low[1,2]	High[1]				[8]
fastMNN	PCA	MNN	MNN–GK	G		High[2]		Balanced		[8]
Seurat v2 (CCA)	CCA	DTW	DTW	E		High[2]				[9]
Seurat v3	CCA	MNN	MNN–GK	G	Low[1,2]	High[2]		Batch correction	High[2]	[10]
Scanorama	SVD	MNN	MNN–GK	G / E	High[2]	Low[2,11]		Balanced	High[1]	[11]
BBKNN	PCA	kNN	-	W	High[1,2]	Low[1,2]		Batch correction		[12]
Conos	PCA	MNN, kNN	-	W				Bio-conservation		[13]
Harmony	PCA	Maximum diversity clustering	Linear mixture model	E	High[2]				High[2]	[14]
DESC	AE	Louvain clustering	Model fine tuning	E			Yes	Bio-conservation		[15]
LIGER	iNMF	Louvain clustering	Quantile normalization	E	Low[2]	Low[2]		Batch correction	High[2]	[16]
scMerge	PCA[c]	MNC	Unwanted variation removal	G	Low[2]	High[2]				[4]
scVI	VAE	-	-	E		Low[1]	Yes	Balanced	High[1]	[17]
scANVI	VAE	-	-	E		Low[1]	Yes	Bio-conservation	High[1]	[18]
scGen	VAE	-	-	G	Low[1,2,e]		Yes	Bio-conservation	High[1]	[19]
trVAE	VAE	-	-	E	Low[1,e]	High[1]	Yes			[20]
SAUCIE	AE	-	MMD regularization	G / E	High[1]		Yes	Batch correction		[21]

PCA, principal component analysis; CCA, canonical correlation analysis; SVD, singular value decomposition; AE, autoencoder; iNMF, integrative nonnegative matrix factorization; VAE, variational autoencoder; MNN, mutual nearest neighbors; DTW, dynamic time warping; kNN, k-nearest neighbors; LD, linear decomposition; GK, gaussian kernel.

[a]The default dimension reduction methods used are shown. [b]ZINB-WaVE performs simultaneously batch effect correction and dimension reduction. [c]scMerge uses both HVG expression space and PCA embeddings. [d]G, Gene expression matrix; E, Embeddings; W, (Weighted edged) Graph. [e]scGen, trVAE were tested on a CPU, while they are optimized for GPU.

Molecules and Cells

REFERENCES

1. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41-50.

2. Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. *21*, 12.

3. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

4. Lin, Y., Ghazanfar, S., Wang, K.Y.X., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.G., Ormerod, J.T., Speed, T.P., Yang, P., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. Proc. Natl. Acad. Sci. U. S. A. *116*, 9775-9784.

5. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., and Theis, F.J. (2019). A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods *16*, 43-49.

6. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

7. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284.

8. Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421-427.

9. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411-420.

10. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888-1902.e21.

11. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685-691.

12. Polański, K., Young, M.D., Miao, Z., Meyer, K.B., Teichmann, S.A., and Park, J.E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics *36*, 964-965.

13. Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K., and Kharchenko, P.V. (2019). Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nat. Methods *16*, 695-698.

14. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289-1296.

15. Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G., and Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat. Commun. *11*, 2338.

16. Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. Cell *177*, 1873-1887.e17.

17. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053-1058.

18. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol. Syst. Biol. *17*, e9620.

19. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715-721.

20. Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36*(Suppl_2), i610-i617.

21. Amodio, M., van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., et al. (2019). Exploring single-cell data with deep multitasking neural networks. Nat. Methods *16*, 1139-1145.