nature protocols

https://doi.org/10.1038/s41596-024-00991-3

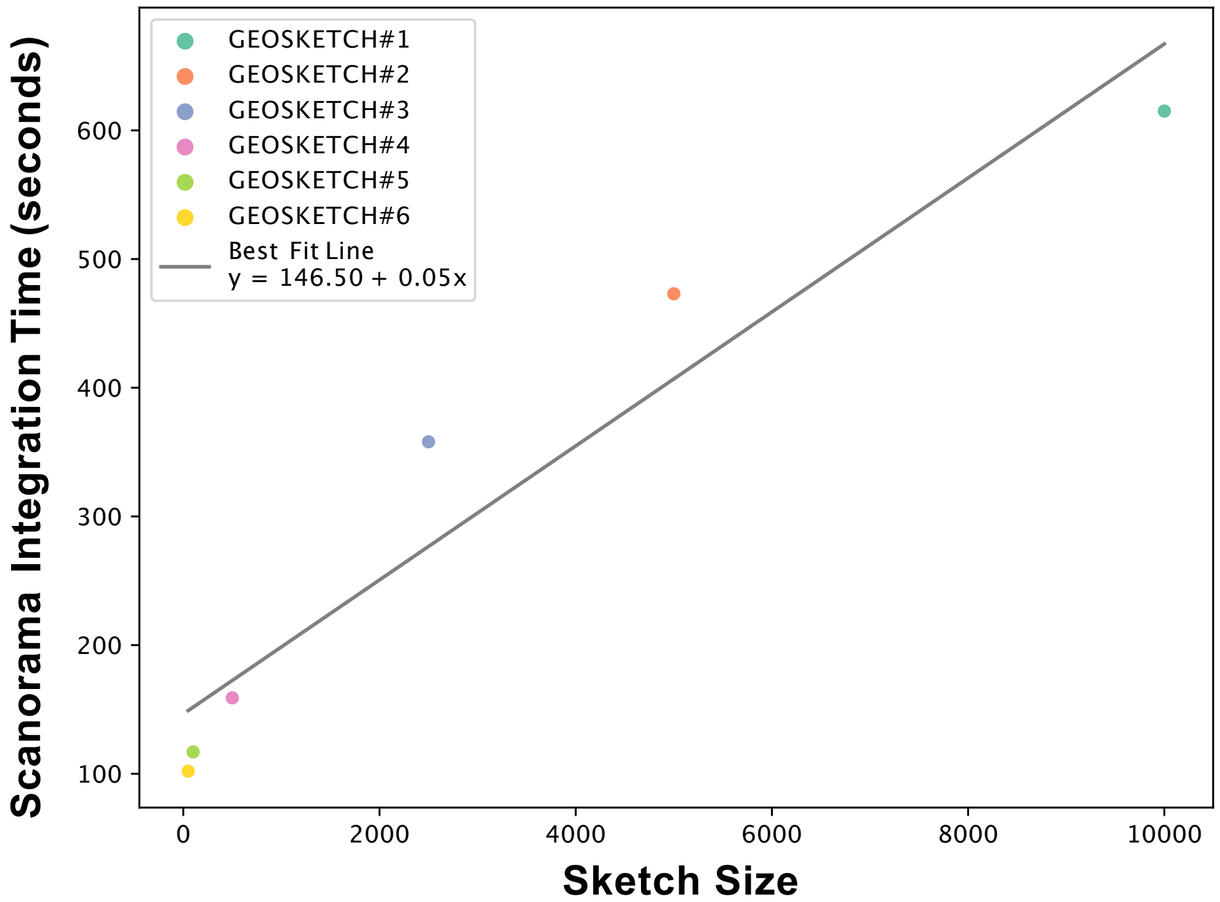
Protocol

Check for updates

Scanorama: integrating large and diverse single-cell transcriptomic datasets

In the format provided by the authors and unedited

a



b

		Bio conservation					Batch correction					Aggregate score		
Geosketch	**Method**	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
NA	**Scanorama**	0.68	0.72	0.52	0.60	1.00	0.77	0.02	0.26	0.62	0.07	0.35	0.70	0.56
10000	**GEOSKETCH #1**	0.67	0.70	0.47	0.60	1.00	0.79	0.02	0.26	0.62	0.09	0.36	0.69	0.56
5000	**GEOSKETCH #2**	0.67	0.70	0.47	0.60	1.00	0.79	0.02	0.25	0.61	0.09	0.36	0.69	0.55
2500	**GEOSKETCH #3**	0.67	0.70	0.47	0.60	1.00	0.79	0.02	0.25	0.62	0.09	0.36	0.69	0.55
500	**GEOSKETCH #4**	0.57	0.67	0.44	0.58	1.00	0.86	0.02	0.24	0.61	0.13	0.37	0.65	0.54
NA	**Unintegrated**	0.72	0.73	0.51	0.63	1.00	0.58	0.01	0.10	0.65	0.00	0.27	0.72	0.54
100	**GEOSKETCH #5**	0.64	0.69	0.47	0.55	1.00	0.80	0.02	0.25	0.61	0.02	0.34	0.67	0.54
50	**GEOSKETCH #6**	0.64	0.68	0.45	0.55	1.00	0.80	0.02	0.22	0.61	0.03	0.33	0.66	0.53

Supplementary Figure 1. Overview of top and bottom ranked sketch parameter adjustments by aggregate score for the large dataset 1 example. **(a)** Scatterplot of sketch size (ranging from 50 to 10,000) and Scanorama integration time in seconds. **(b)** Single-cell integration benchmarking (Scib) metrics was applied to assess the quality of the Scanorama-integrated 26 single-cell datasets with a range of sketching subsamples. Metrics are divided into bio conservation and batch correction categories. Aggregate scores are computed using a 40/60 weighted mean of the category scores as done by Luecken et al.[3] The color code represents the range of values for each metric. Circles accompanying metric values are color-coded using a Purple-Green colormap (PRGn). The colormap spans a range of 2.5 standard deviations from the mean, where higher values are depicted in shades of green, and lower values in shades of purple. For the "Total" aggregate score, represented by bars in the figure, a distinct color scheme is applied. The color of the bars is determined by the Yellow-Green-Blue color map (YIGnBu). This color code reflects a weighted combination of "Batch correction" and "Bio conservation" scores. Higher "Total" scores are depicted shades of blue, while lower scores are in shades of yellow-green.

				Bio conservation					Batch correction					Aggregate score		
Knn	Sigma	Dimred	Method	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
20	1	100	Scanorama_sig	0.68	0.72	0.51	0.60	1.00	0.77	0.02	0.26	0.62	0.09	0.35	0.70	0.56
20	5	100	Scanorama_sig2	0.68	0.72	0.51	0.60	1.00	0.77	0.02	0.26	0.62	0.08	0.35	0.70	0.56
20	15	100	Scanorama	0.68	0.72	0.52	0.60	1.00	0.77	0.02	0.26	0.62	0.07	0.35	0.70	0.56
30	15	100	Scanorama_knn3	0.67	0.69	0.50	0.60	1.00	0.79	0.02	0.26	0.62	0.07	0.35	0.69	0.56
20	50	100	Scanorama_sig3	0.68	0.72	0.53	0.60	1.00	0.76	0.03	0.26	0.59	0.03	0.33	0.70	0.56
40	15	100	Scanorama_knn4	0.68	0.69	0.50	0.60	1.00	0.78	0.02	0.26	0.62	0.05	0.35	0.69	0.55
50	15	100	Scanorama_knn5	0.68	0.69	0.50	0.60	1.00	0.79	0.02	0.27	0.61	0.05	0.35	0.69	0.55
20	15	50	Scanorama_dim2	0.70	0.71	0.50	0.62	1.00	0.70	0.02	0.23	0.69	0.00	0.33	0.70	0.55
10	15	100	Scanorama_knn2	0.68	0.72	0.48	0.60	1.00	0.75	0.02	0.25	0.60	0.07	0.34	0.70	0.55
5	15	100	Scanorama_knn	0.68	0.72	0.53	0.60	1.00	0.69	0.02	0.14	0.56	0.08	0.30	0.71	0.54
20	15	10	Scanorama_dim	0.74	0.68	0.48	0.63	0.99	0.55	0.02	0.20	0.69	0.00	0.29	0.70	0.54
NA	NA	NA	Unintegrated	0.72	0.73	0.51	0.63	1.00	0.58	0.01	0.10	0.65	0.00	0.27	0.72	0.54
20	100	100	Scanorama_sig4	0.67	0.73	0.50	0.60	1.00	0.75	0.03	0.23	0.46	0.01	0.29	0.70	0.54
20	15	1000	Scanorama_dim4	0.61	0.76	0.52	0.54	1.00	0.88	0.03	0.14	0.08	0.44	0.31	0.68	0.54
20	15	500	Scanorama_dim3	0.63	0.73	0.49	0.56	1.00	0.84	0.03	0.19	0.17	0.33	0.31	0.68	0.53

Supplementary Figure 2. Overview of top and bottom ranked parameter adjustments by aggregate score for the large dataset 1 example.
Single-cell integration benchmarking (Scib) metrics was applied to assess the quality of the Scanorama-integrated 26 single-cell datasets from 9 different technologies. Metrics are divided into bio conservation and batch correction categories. Aggregate scores are computed using a 40/60 weighted mean of the category scores as done by Luecken et al.[3] The color code represents the range of values for each metric. Circles accompanying metric values are color-coded using a Purple-Green colormap (PRGn). The colormap spans a range of 2.5 standard deviations from the mean, where higher values are depicted in shades of green, and lower values in shades of purple. For the "Total" aggregate score, represented by bars in the figure, a distinct color scheme is applied. The color of the bars is determined by the Yellow-Green-Blue color map (YlGnBu). This color code reflects a weighted combination of "Batch correction" and "Bio conservation" scores. Higher "Total" scores are depicted shades of blue, while lower scores are in shades of yellow-green.

Benchmarking Small Dataset

	Bio conservation					Batch correction					Aggregate score		
Method	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
Harmony	0.78	0.96	0.98	0.78	1.00	0.98	0.36	0.73	0.92	0.02	0.60	0.90	0.78
scANVI	0.94	0.98	0.99	0.93	1.00	0.82	0.07	0.32	0.99	0.00	0.44	0.97	0.76
Scanorama	0.73	0.96	0.98	0.73	1.00	0.97	0.35	0.60	0.69	0.19	0.56	0.88	0.75
Unintegrated	0.77	0.96	0.98	0.77	1.00	0.89	0.05	0.31	0.92	0.00	0.44	0.90	0.71
scVI	0.58	0.18	0.11	0.57	0.97	0.72	0.13	0.10	0.93	0.92	0.56	0.48	0.51

Benchmarking Large Dataset

	Bio conservation					Batch correction					Aggregate score		
Method	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Total
scANVI	0.84	0.84	0.68	0.76	1.00	0.77	0.01	0.26	0.90	0.30	0.45	0.82	0.67
Scanorama	0.68	0.72	0.52	0.60	1.00	0.77	0.02	0.26	0.62	0.07	0.35	0.70	0.56
Unintegrated	0.72	0.73	0.51	0.63	1.00	0.58	0.01	0.10	0.65	0.00	0.27	0.72	0.54
Harmony	0.47	0.48	0.20	0.47	0.98	0.83	0.04	0.21	0.63	0.11	0.36	0.52	0.46
scVI	0.28	0.34	0.15	0.35	0.99	0.55	0.02	0.21	0.64	0.97	0.48	0.42	0.44

Supplementary Figure 3. Benchmarking results for the small and large dataset.
Single-cell integration benchmarking (Scib) metrics was applied to assess the quality of the superiority of Scanorama integration method against other integration methods, scANVI, scVI and Harmony. Metrics are divided into bio conservation and batch correction categories. Aggregate scores are computed using a 40/60 weighted mean of the category scores as done by Luecken et al.[3] The color code represents the range of values for each metric. Circles accompanying metric values are color-coded using a Purple-Green colormap (PRGn). The colormap spans a range of 2.5 standard deviations from the mean, where higher values are depicted in shades of green, and lower values in shades of purple. For the "Total" aggregate score, represented by bars in the figure, a distinct color scheme is applied. The color of the bars is determined by the Yellow-Green-Blue color map (YlGnBu). This color code reflects a weighted combination of "Batch correction" and "Bio conservation" scores. Higher "Total" scores are depicted shades of blue, while lower scores are in shades of yellow-green.