Supplementary information

Index

Table S1. Dataset availability. Links to the datasets and PubMed Identifiers of the related scientific papers

Dataset - name of the dataset.
Link - the resource where the dataset is published.
Registration Required - indicator showing whether dataset access requires user registration: “No” indicates that the dataset can be downloaded directly via the provided link without registration, while “Yes” indicates that registration is necessary, although no special request is required.
PMID - PubMed Identifier of the related scientific paper.

Table S2. Sample metadata. Clinical information for all samples, including tumor characteristics, receptor status, Ki67 data, and assigned clinical subtype

Sample ID – unique identifier assigned to each sample, as reported in the original study.
Dataset ID – identifier of the dataset the sample is associated with.
TNM – Tumor-Node-Metastasis (TNM) classification of the sample.
Stage – clinical stage of the cancer.
Sample Type – origin type of the sample.
Sex – sex of the patient the sample was obtained from.
Age – age of the patient the sample was obtained from.
ER Status - Estrogen Receptor expression status: positive or negative.
PR Status - Progesteron Receptor expression status: positive or negative.
HER2 Status - Human Epidermal Growth Factor Receptor 2 expression status: positive or negative.
Ki67 Percentage - percentage of tumor cells exhibiting Ki-67 immunoreactivity.
Clinical Subtype - subtype determined based on hormone receptor and HER2 expression statuses.
Patient ID – internal identifier used to distinguish individual patients.

Table S3. Summary of dataset technologies and per-dataset statistics following quality control. Reference genome version, sequencing technology, DNA library, number of samples after quality control and number of cells after each quality control step

Dataset Name - name of the dataset.
Reference Genome - version of the reference human genome.
Sequencing Technology - technology applied while sample sequencing.
DNA Library - scRNA-seq library preparation chemistry used in the 10x Genomics Chromium platform.
Number of Samples - number of samples after the quality control (QC) pipeline.
Cells Initial - initial number of cells before QC.
Cells Min Genes - number of cells retained after filtering out cells with < 100 genes expressed.
Cells MT Content - number of cells retained after filtering out cells with mitochondrial content > 10%.
Cells Max Genes - number of cells retained after filtering out cells with > 2500 (> 7000) genes expressed.
Cells 2SD - number of cells retained after filtering out cells with $\log_{10}$-transformed unique molecular identifier counts lying outside 2 standard deviations from the mean.

Table S4. Differentially expressed genes across clusters. List of genes showing significant expression differences between clusters. For each gene, statistical significance, fold change, fraction of expressing cells, and the target cluster are provided.

Gene - gene name.
p-value - raw p-value from the Wilcoxon rank-sum test.
Average log2FC - the average log2 fold change in expression between the target cluster and the reference group (all other clusters).
pct.1 - the fraction of cells in the target cluster expressing the gene.
pct.2 - the fraction of cells in the reference group expressing the gene.
Adjusted p-value - adjusted p-value after multiple testing correction using the Bonferroni method.
Cluster - ID of the target cluster.

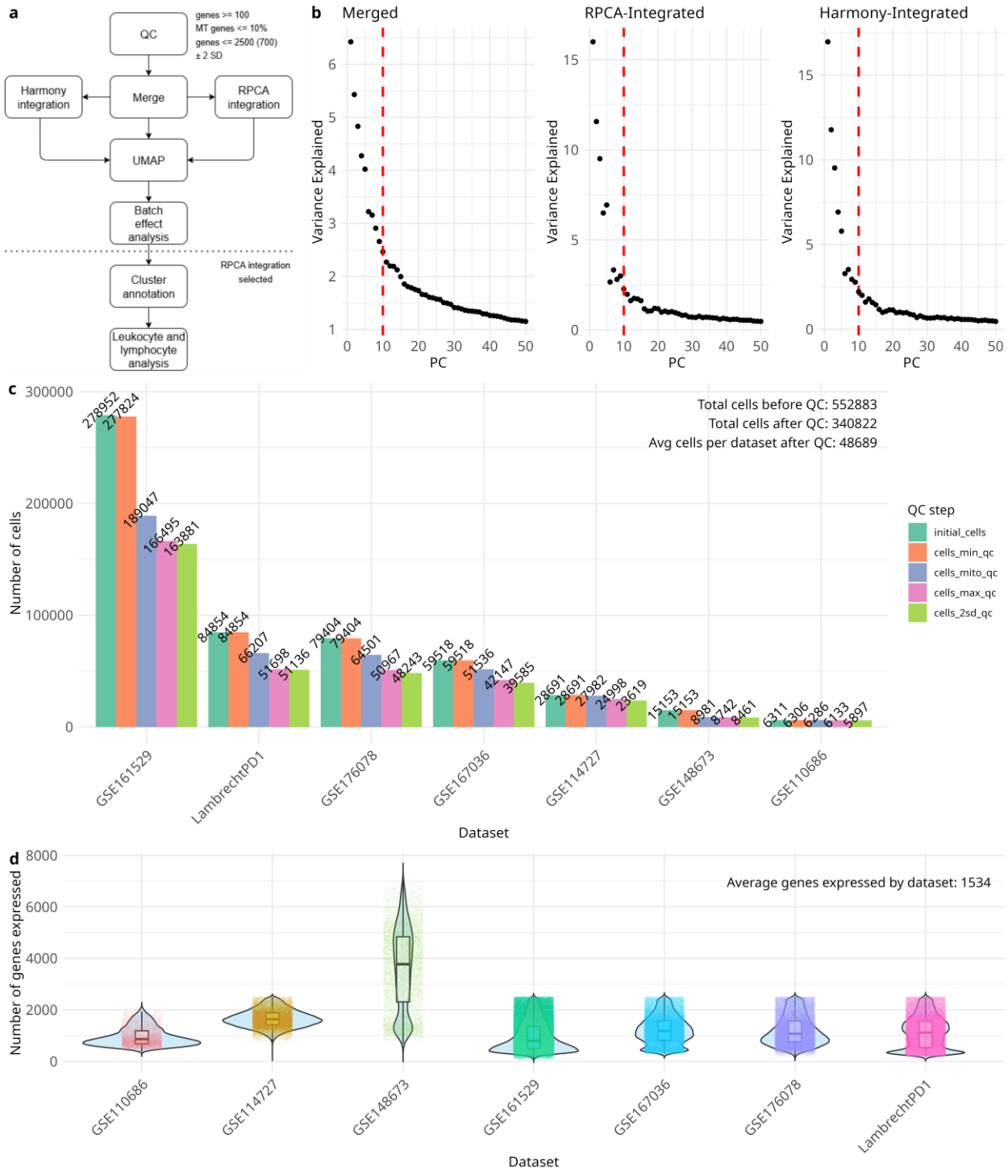

Figure S1. Overview of methods, quality control statistics, and dimensionality selection for clustering. a Diagram with the overview of methods. **b** Elbow plots for the merged, RPCA-integrated, and Harmony-integrated objects. The y-axis shows the variance explained by each principal component, and the vertical red dashed line indicates the number of dimensions selected for downstream analyses. **c** Number of cells retained in each dataset after each quality control (QC) step. **d** Violin plot representing the distribution of the number of expressed genes per cell across dataset, as a measure of RNA quality. All violin plots indicate median (center line), 25th and 75th percentiles (bounds of box), and minimum and maximum (whiskers). Each dot represents a cell.

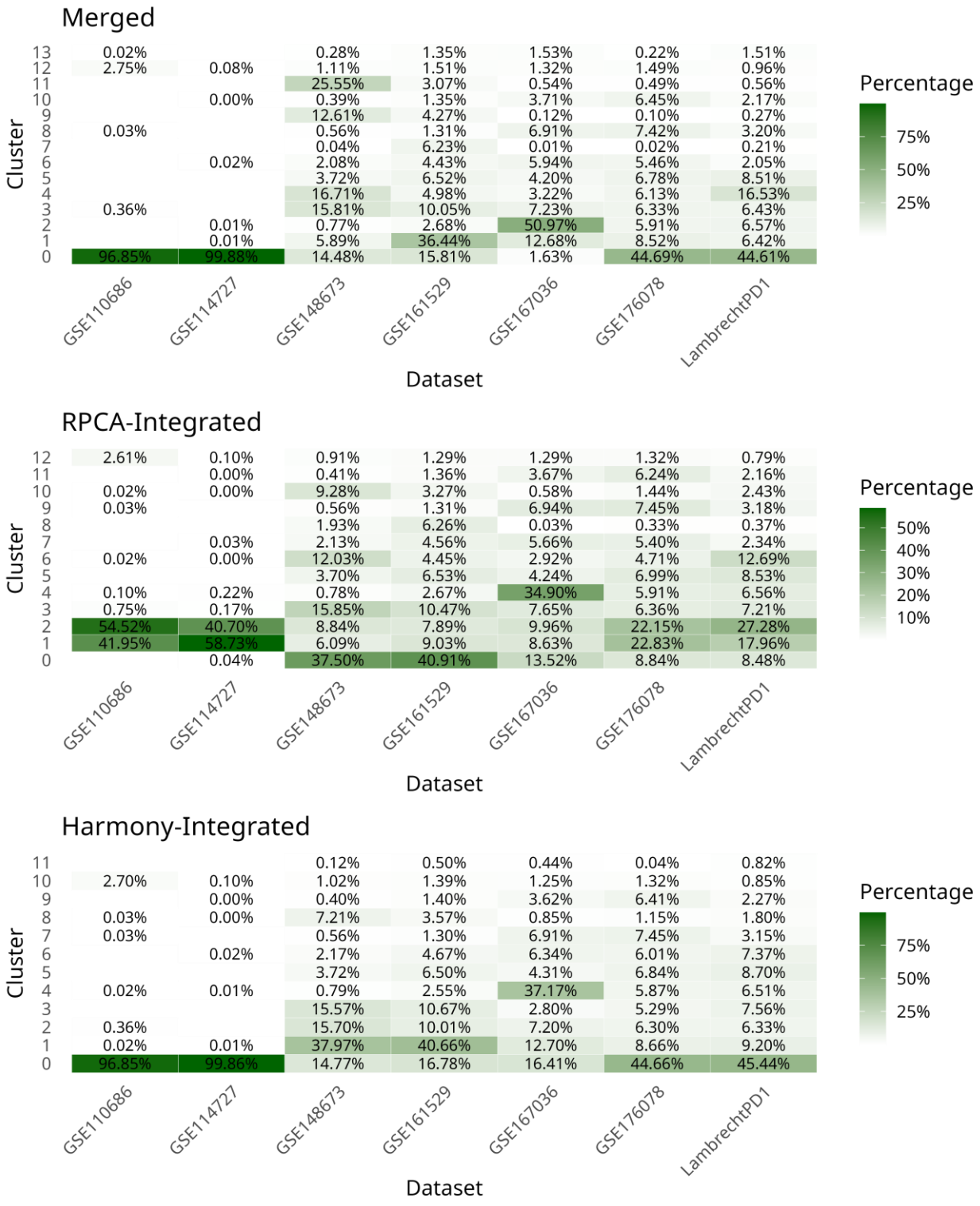

Figure S2. Dataset distribution across clusters for the merged, RPCA-integrated and Harmony-integrated objects. Percentages on the bars indicate the proportion of each dataset that belongs to a specific cluster.