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 ~1 million 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.