Summary of input data

Cellular gene expression profiles were generated from the developing organ, an embryonic tissue that gives rise to organs, of wild-type species. Four biological replicates were collected, and the 10x Genomics scRNA-seq protocol was used to generate transcriptomic profiles at the single-cell level. The scRNA-seq datasets contain gene expression profiles of 4100, 6640, 9090, and 10760 cells with an average median of 5195 genes and an average median of 21,637 unique molecular identifier (UMI) counts per cell. The reference genome of this species contains a total of 27,773 genes.

Quality Control of the scRNA-seq Datasets

The pre-processing of the scRNA-seq datasets involved several steps where each dataset was subjected to filters to prepare them for downstream analyses. First, uninformative genes induced by protoplasting (1180 genes, identified from comparing bulk RNA-seq experiments between protoplast and un-protoplast tissues) were removed. After removing protoplasting-induced genes, genes for which expression was not detected in a minimum of 0.2% of the cell population (1 out of 500 cells) were excluded from the scRNA-seq data. Next, cells expressing less than 200 genes or with counts larger than 110,000 were excluded. Cells were further filtered based on the proportion of mitochondrial (5% or more) and chloroplast (10% or more) gene counts. After removing low-quality cells, the number of cells in each replicate was 4040, 6614, 9007, and 10589, totaling 30,250 cells. After filtering, log-normalization was applied to the data using Seurat to control for variation in sequencing depth and cell efficiency.

To reduce the dimensionality of the data before performing statistical analyses, a set of highly variable genes (HGV) was selected. This was done by calculating gene-specific mean-variance relationships across cells in each replicate and selecting the 2000 genes that are outliers on each replicate's mean-variance relationship. In the next step, 2000 top-scoring highly variable genes were selected by ranking the variable genes based on the number of datasets in which they were deemed variable. Finally, gene-level unit-variance scaling was performed to ensure that each gene had the same variance, and scaling was done before implementing factorization and integration using non-negative matrix factorization (Liger). We do not center the data by subtracting the mean because non-negative matrix factorization requires positive values.

Representation of the combined replicates data

This section presents the method used to represent the integrated or combined replicates data and identify cell clusters and their constituent gene expression programs (GEPs). Our approach involved applying the integrative non-negative matrix factorization method implemented in the software Liger. We aimed to decipher the

molecular makeup of a cell or cell type based on the information provided by highly variable genes that partition the data into biologically meaningful cell types.

To begin, all replicates of the species were used to conduct the matrix factorization method implemented in Liger, integrating the datasets into a shared space. This space, named the coefficient matrix, represents a reduced representation of the combined datasets. Liger produced four types of matrices: a dataset-specific basis matrix and a dataset-specific coefficient matrix for each dataset, a shared basis matrix, and a shared coefficient matrix for all datasets. By utilizing all the outputs from Liger, we could identify sets of co-expressed genes or GEPs shared between replicates and GEPs that are replicate-specific from the shared basis matrix. This information allowed us to estimate the molecular makeup of individual cells or cell types.

The shared coefficient matrix shows a lower-dimensional representation of the integrated datasets in a common feature space. The values in the coefficient matrix represent the proportion of GEP usage by each cell, while the values in the basis matrix show how each GEP is formed based on the relative contribution of each gene to each GEP. We used the Louvain clustering method to identify cell clusters based on the similarity of the molecular composition of the cells in the coefficient matrix.

To factorize the datasets, Liger was run for a range of GEPs (also referred to as components or latent factors or metagenes), K = 10 to 50. Throughout this study, we use the terms latent factor and GEP interchangeably. For the integrative analysis of the species, we selected the decomposition of the datasets with K = 44 GEPs.

Defining the GEPs and their usage by the cells

To obtain a deeper understanding of the gene expression profile of cells and a more detailed view of cellular arrangement during cellular activities, we utilized the NMF-based Liger tool with a range of factorization components, K = 10 to 50, to decompose the combined replicates data. Figure 2A presents a tree of co-expressed genes or GEPs identified from the shared basis matrix for multiple values of decomposition components (K). Each node represents a set of co-expressed genes annotated with enriched GO terms (see Section 2.2.8.7). In Figure 2B, the GO terms retrieved for the genes in GEP-2 (of factorization K = 10) are shown, which are involved in cell-cycle-related cellular activities. Based on the GO annotation of the sets of GEPs, we characterized them as identity GEPs or activity GEPs. As shown in Figure 2A, the GO annotation of the sets of GEPs for factorization K = 10 revealed five identity GEPs and five activity GEPs. This feature of the Liger method and NMF-based methods in general, therefore, allows us to understand the causes underlying the partitioning of cells into different clusters.

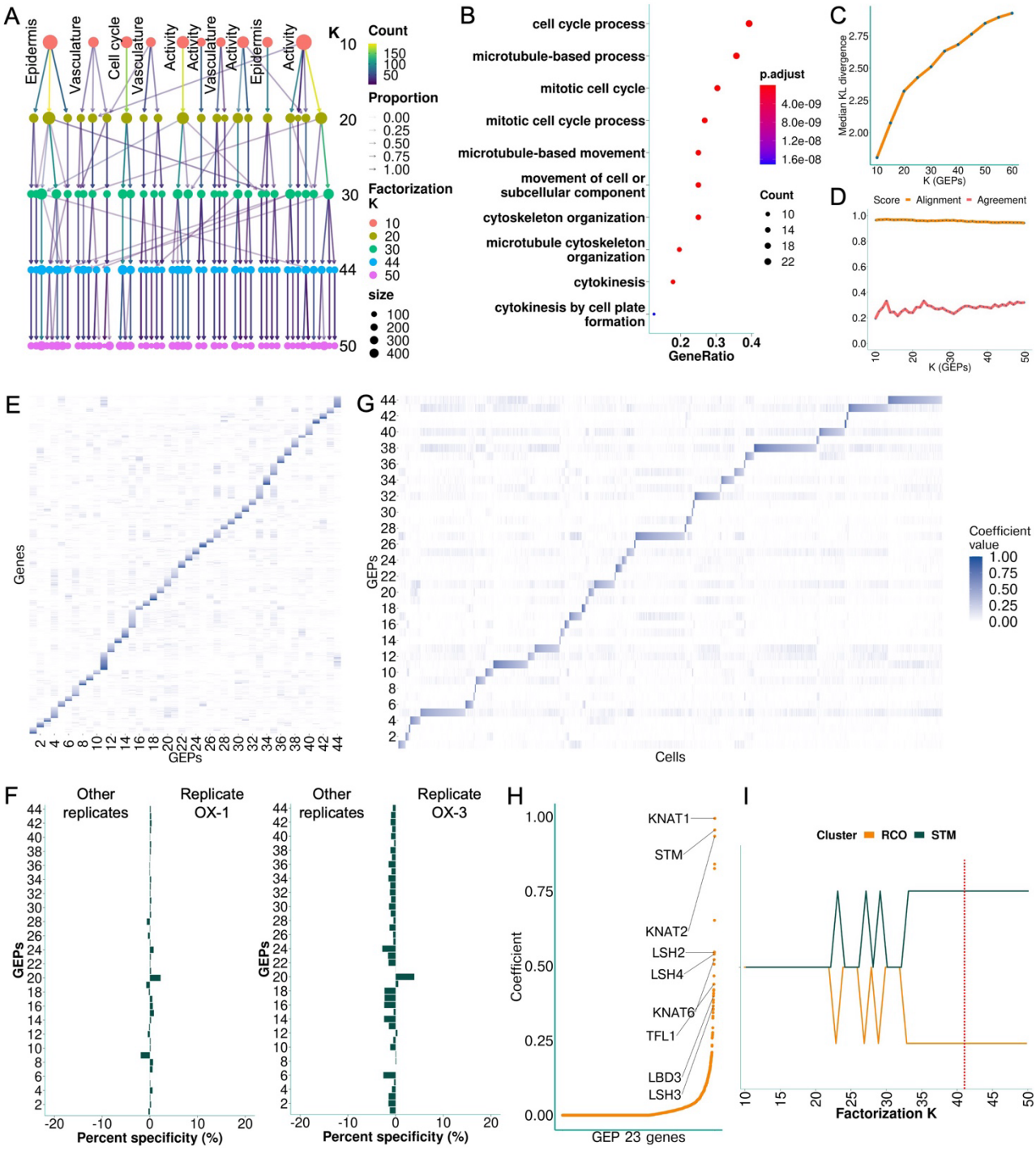

Figure 2: Identification and characterization of GEPs and their usage by the cells. A) Annotated tree of GEPs identified from the shared basis matrix from different factorization *K* (10, 20, 30, 44, and 50). Nodes are annotated with enriched GO terms for *K* = 10 (first row). Node, GEP; color, factorization component (*K*); edge color, the proportion of genes moving from one node to another; node size, number of genes in the node. B) Enriched GO terms retrieved for the genes in GEP-2 of factorization rank K = 10. C) Visual guide for selecting the number of components, K. Median Kullback-Leibler (K-L) divergence from uniform for cell factor loadings is plotted as a function of K. D) Alignment and agreement score. The alignment score quantifies how well-aligned two or more datasets are; the agreement score quantifies how much the factorization and alignment distort the geometry of the original datasets. E) Heatmap of the shared basis matrix. Each column represents a GEP, and each row represents the relative expression of genes to the GEPs. Dark blue indicates higher expression of genes to the GEP. F) Heatmap of the shared coefficient matrix. Each column represents a cell, and each row represents.

the proportion of GEP used by the cells. Dark blue indicates a higher percentage of GEP usage. G) Specificity of GEPs to the replicates. Positive values indicate specificity to a specific replicate, while negative values indicate specificity to the rest of the replicates. Specificity score has been shown for OX-1 (left) and OX-3 (right) replicates. H) Top most significant genes in GEP-23. Only those were shown for which one-to-one orthologues were present. I) Clustering of the majority of RCO- and STM-expressing cells, obtained from the graph-based clustering method (resolution parameter, r = 0.3), applied to the coefficient matrix for a range of factorization K = 10 to 50. Yellow, RCO-expressing cells; green, STM-expressing cells. Line overlaps when the majority of the cells expressing these two genes fall in the same cluster, and a split indicates they fall in different clusters. The red line indicates the smallest value of K (41) for which these two cell populations are separated for any value of resolution parameter, r.

Selecting the number of components is one of the most critical steps in analyzing scRNA-seq data. Therefore, we relied on multiple metrics to guide us in selecting an appropriate value of K (see Section 2.2.8.4). Figure 2C shows the median Kullback-Leibler (K-L) divergence from uniform for cell factor loadings as a function of K. We used two other metrics, the alignment score and agreement score, to select an appropriate value of K, as shown in Figure 2D. However, none of these metrics alone were sufficient in suggesting an appropriate value of K, so we relied on the biological interpretation of the GEPs and the arrangement of cell clusters for a particular choice of K. Based on the observation in Figure 2C, we could infer that the appropriate number of components K lies between 40 to 60. Therefore, for the integrative analysis of replicates, we selected factorization rank K = 44 based on the interpretability of GEPs and because it facilitated the description of specific cell types relevant to the biological questions associated with this project.