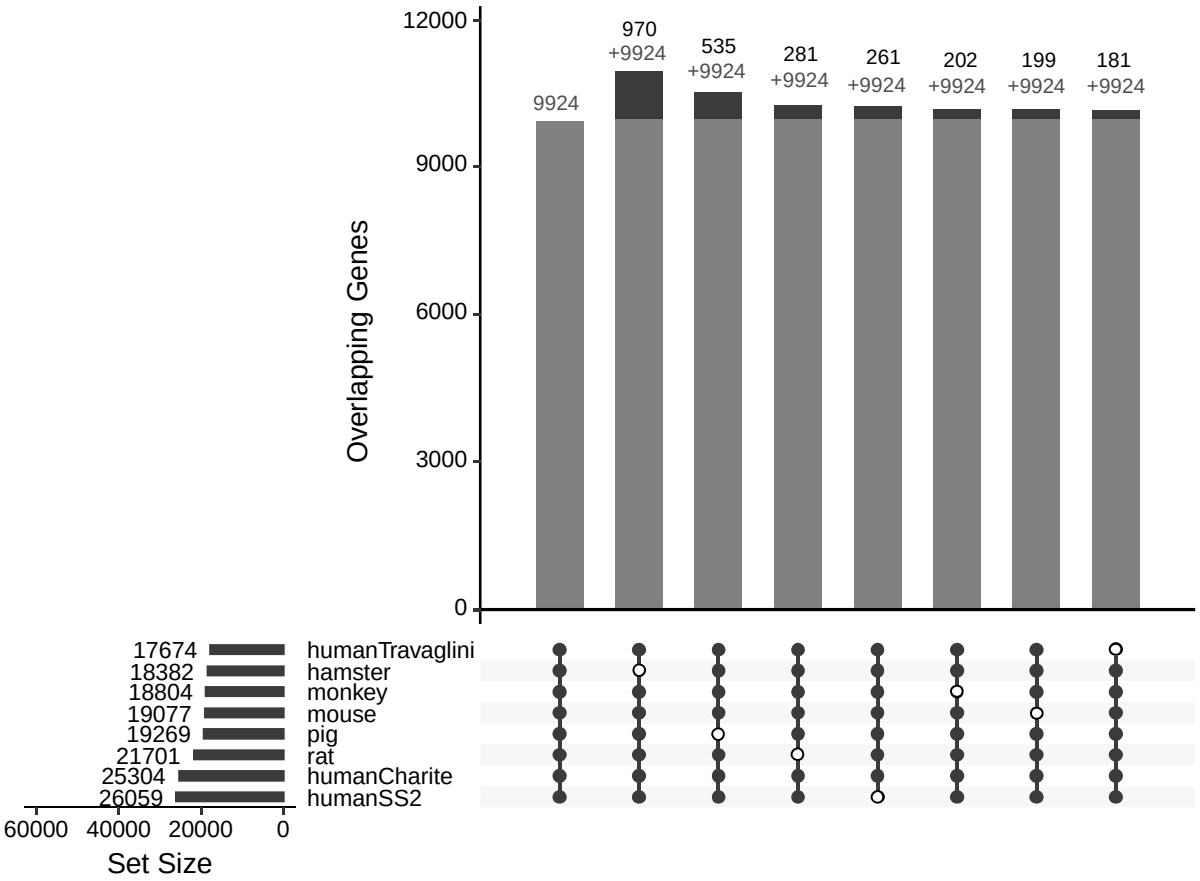
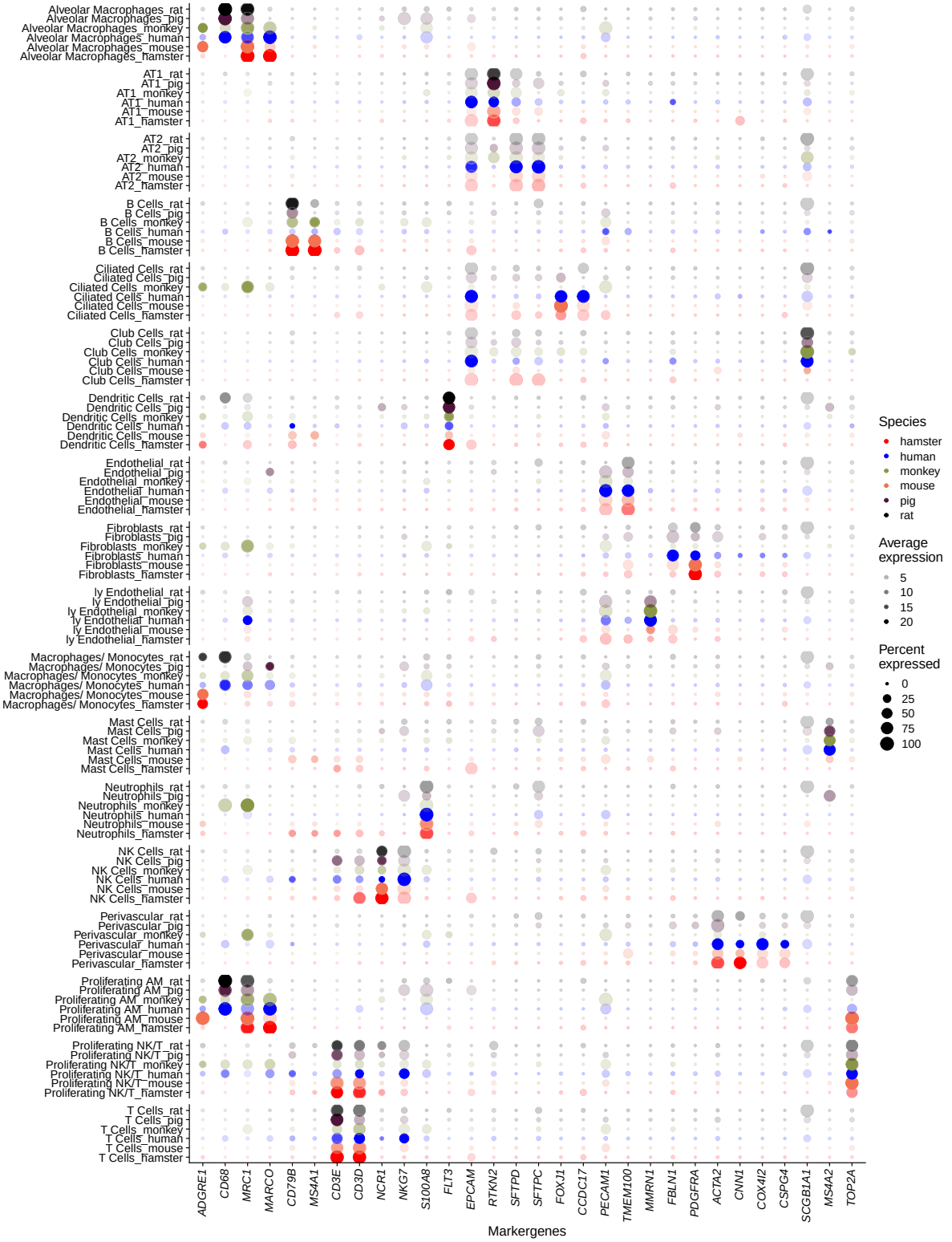
Supplementary material to:

A pulmonologist's guide to perform and analyse cross-species single-lung-cell transcriptomics

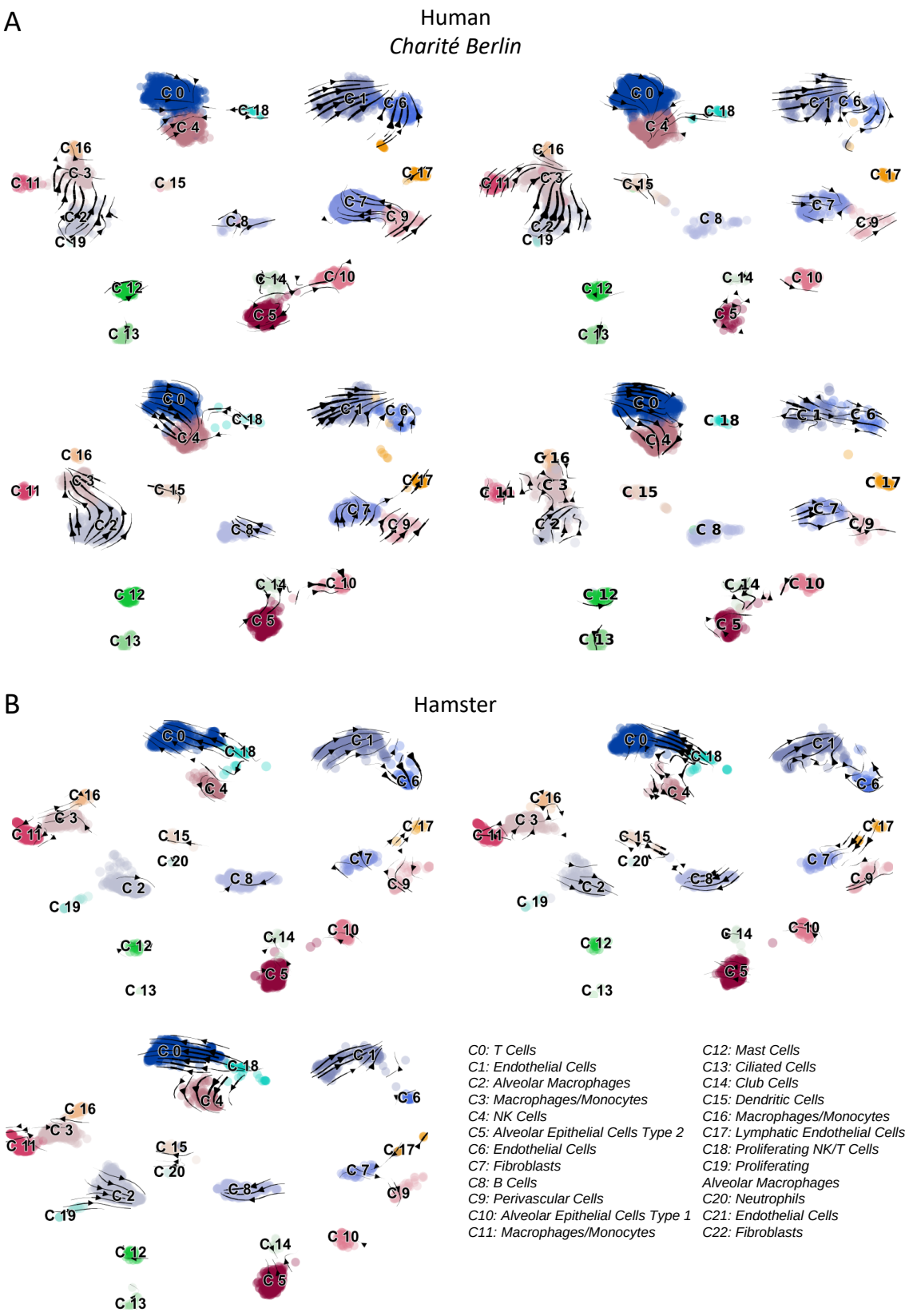
Peter Pennitz[1,2]*, Holger Kirsten[3]*, Vincent D. Friedrich[3,4], Emanuel Wyler[5], Cengiz Goekeri[1,2,6], Benedikt Obermayer[7], Gitta A. Heinz[8], Mir-Farzin Mashreghi[8,9], Maren Büttner[10,11], Jakob Trimpert[12], Markus Landthaler[5,13], Norbert Suttorp[2], Andreas C. Hocke[1,2], Stefan Hippenstiel[2], Mario Tönnies[14], Markus Scholz[3], Wolfgang M. Kuebler[15,16], Martin Witzenrath[1,2,16], Katja Hoenzke[1,2] and Geraldine Nouailles[1,2,#]



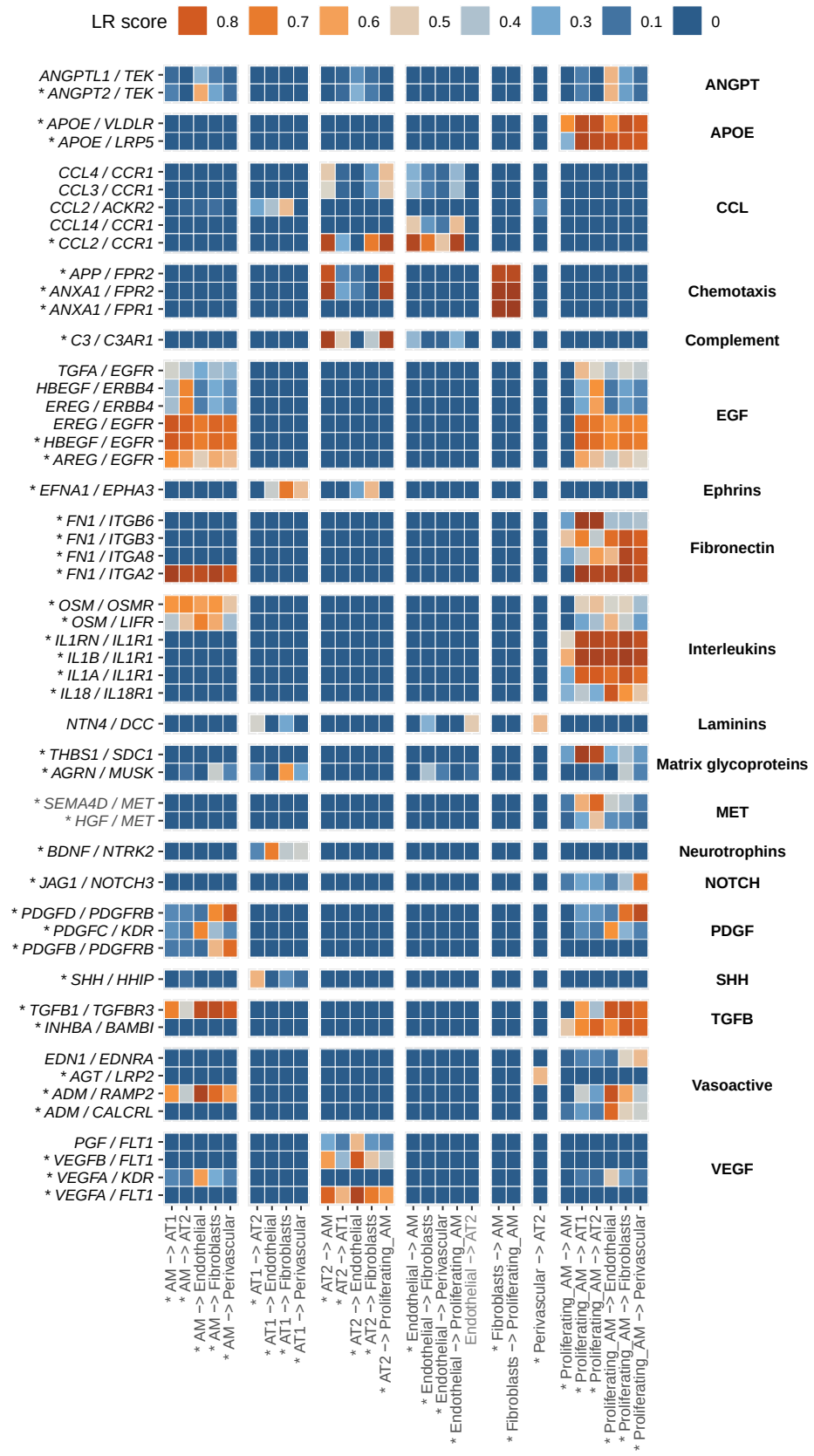
Supplementary Figure 1: Overlapping genes of interspecies datasets**.** For non-human datasets, human orthologues from Ensemble GRCh38.p13 were used. The number to the left labelled as “Set Size” corresponds to the number of unique transcripts in each dataset. The number of genes found in multiple species are indicated by connected dots, e.g. 9924 unique transcripts are found in all species. Without integration of hamster datasets 970 transcripts more would overlap, without pig 535 transcripts, without rat 281 transcripts, without the human SS2 dataset 261 transcripts, without monkey 261 transcripts 202, without mouse datasets 199 transcripts, and without the human Travaligni dataset 181 transcripts.



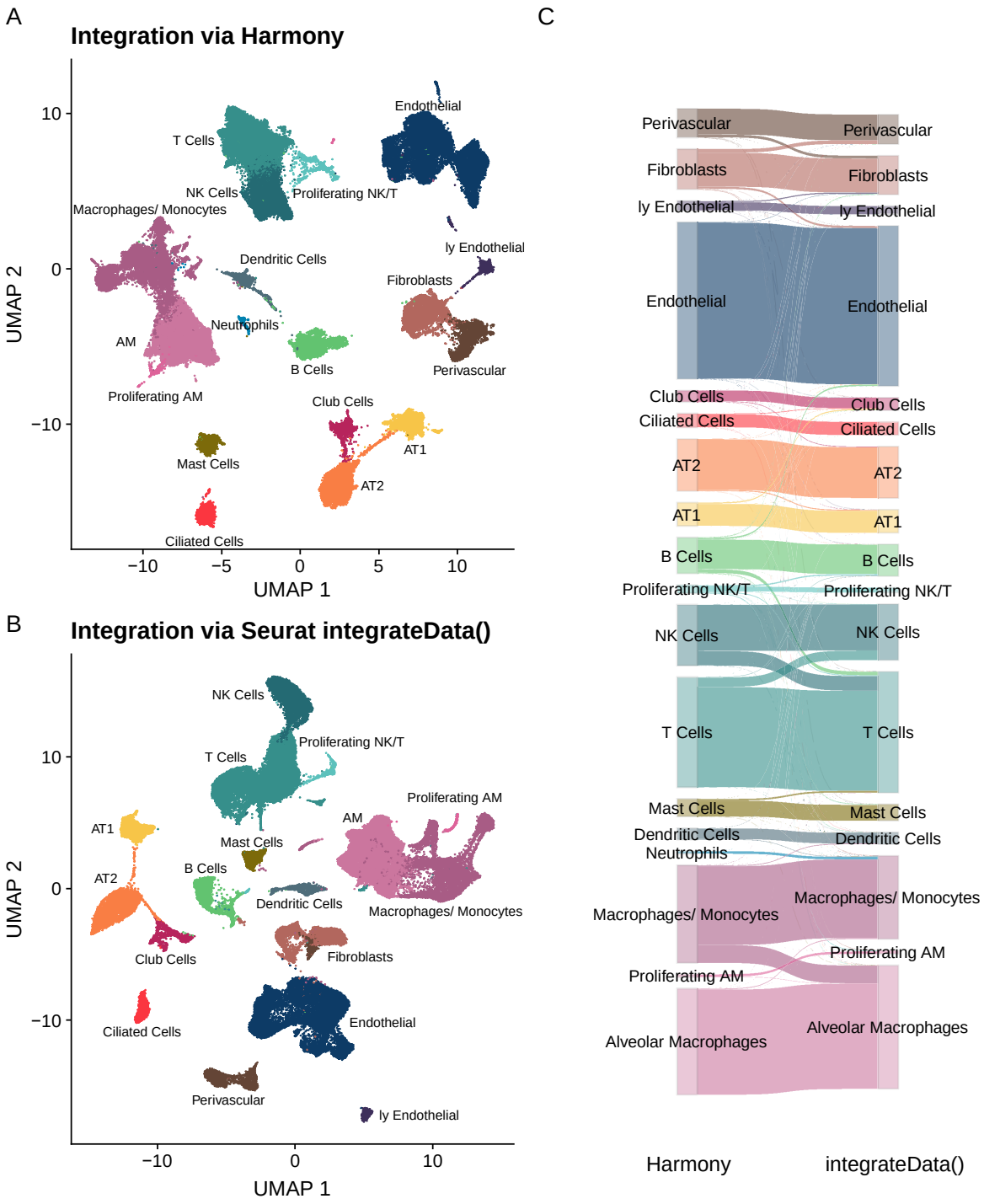
Supplementary Figure 2: Overview of top marker genes used for cell type annotation and their interspecies expression pattern. Dot plot of top mRNA marker genes for interspecies cross-comparison. Point size indicates percentage of cells in a certain population expressing the designated marker gene. Colour type indicates species while colour intensity indicates average expression levels (arbitrary units).



Supplementary Figure 3: RNA velocity of individual batches. UMAP plots depicting individual (A) human and (B) hamster lung cells overlaid with RNA velocity stream arrows. The colouration indicates chosen cell clusters. Arrows indicate the estimated cell state trajectories of cells.



Supplementary Figure 4: Intercellular communication estimation via ligand-receptor co-expression in the designated cell types. Relevant ligand-receptor interactions (LRscore ≥0.5) observed in humans. Human ligand-receptor pairs found to be conserved (i.e. detected in humans and at least one additional non-human species) are indicated with an asterisk and shown in black. Interactions are limited to those classified by Raredon et al. [53].



Supplementary Figure 5: Exemplarily comparison of an alternative integration via (A) package Harmony vs. (B) package Seurat's integrateData() function. Beside the integration step, all other parameter and pre-processing steps remained the same (see Figure 1). Shown is the first and second Dimension of the UMAP projection coloured by assigned cell types. (C) Sankey diagram demonstrating that cell types of clusters obtained in the integrateData plot were assigned according to the cluster-wise majority celltype labels from the harmony-based integration.