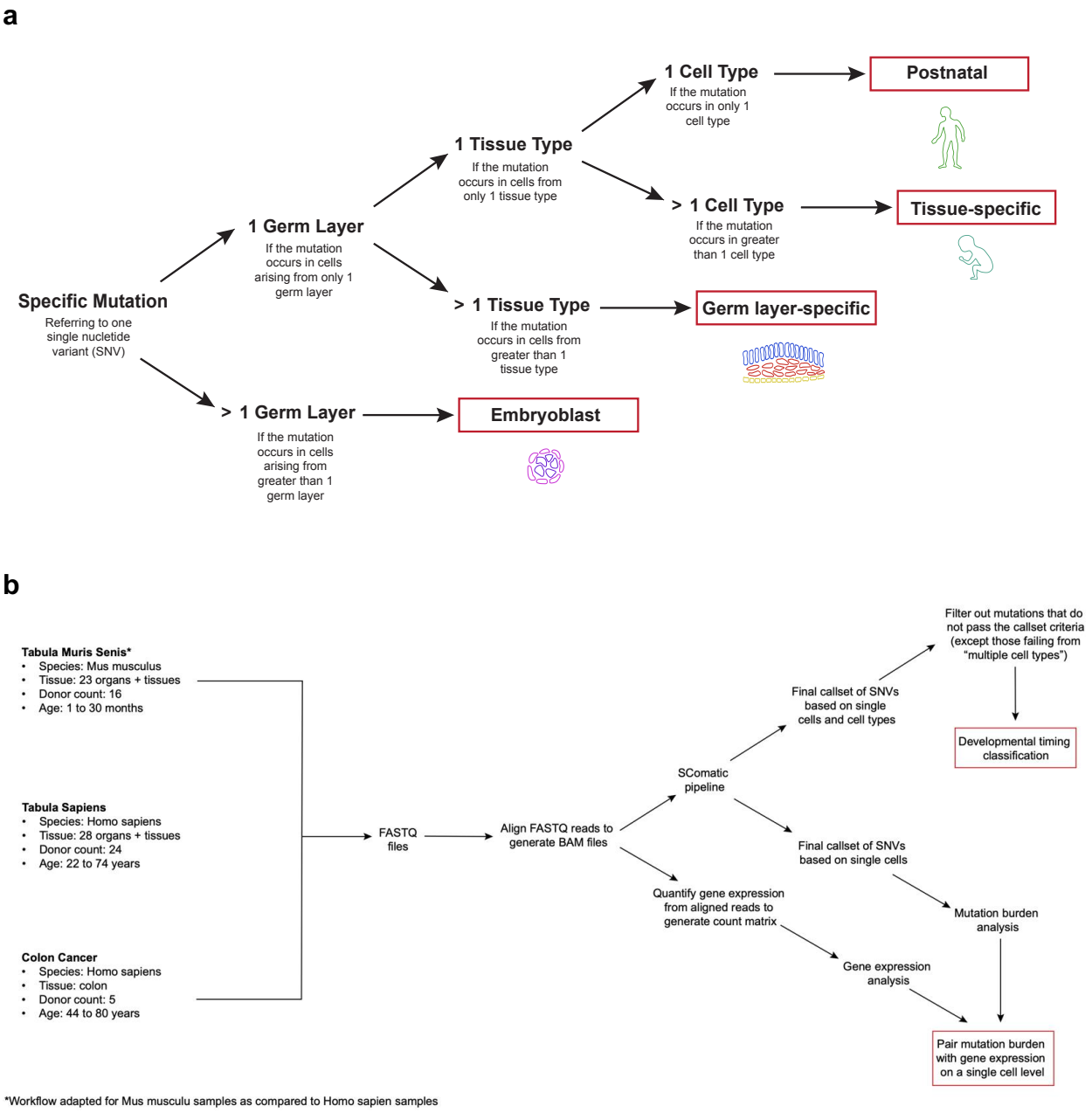
Extended Data:
Fetal Life is the Predominant Source of Somatic Mosaicism Across Tissues and Aging

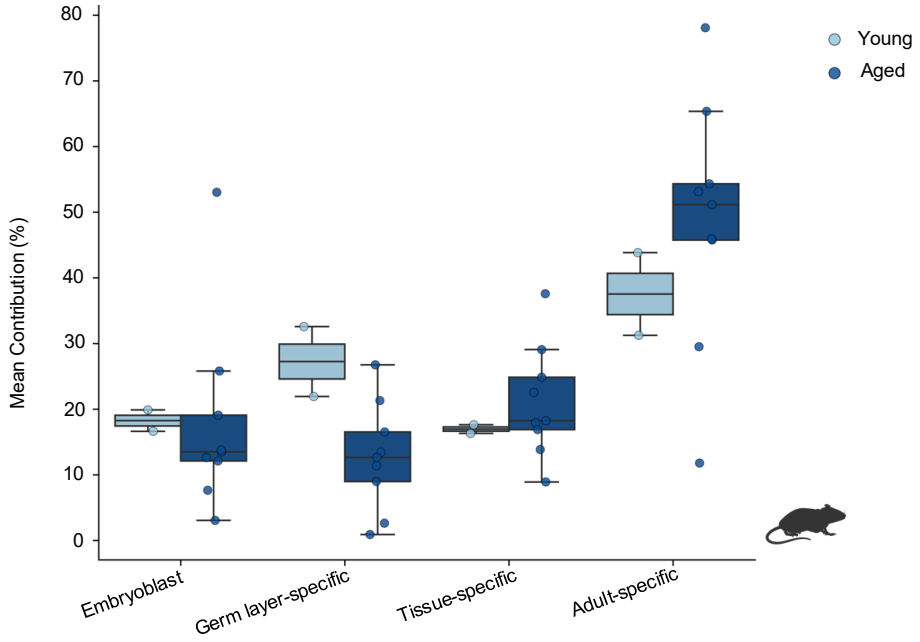
Sierra Lore[1,2], Zane Koch[3], Björn Schumacher[4,5], Trey Ideker[3], Morten Scheibye-Knudsen[2*], Eric Verdin[1*]

*Corresponding author(s). Email: mscheibye@sund.ku.dk, everdin@buckinstitute.org.

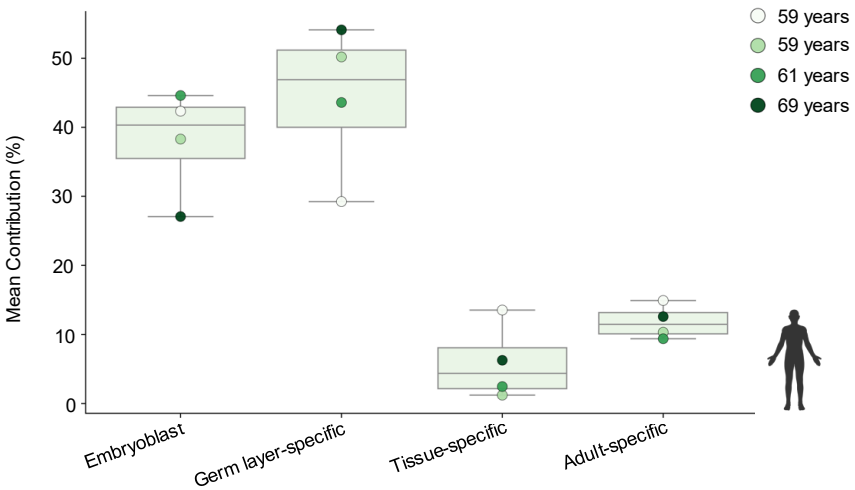


Extended Data Figure 1 | Developmental classification of somatic mutations and integrated single-cell analysis workflow. a, Framework schematic used to assign variants to four developmental timing categories based on their distribution across germ layers, tissues, and cell types. **b,** Overview of the single-cell variant calling and expression-integration pipeline applied to mouse, human, and colon cancer datasets. FASTQ files were aligned to generate BAM files, followed by variant calling using the SComatic workflow to obtain per-cell SNV callsets. After filtering and developmental classification, mutations were quantified for downstream burden analyses. Gene expression profiles derived from aligned reads were paired with SNV calls, enabling joint analysis of mutation burden and transcriptional state at the single-cell level.

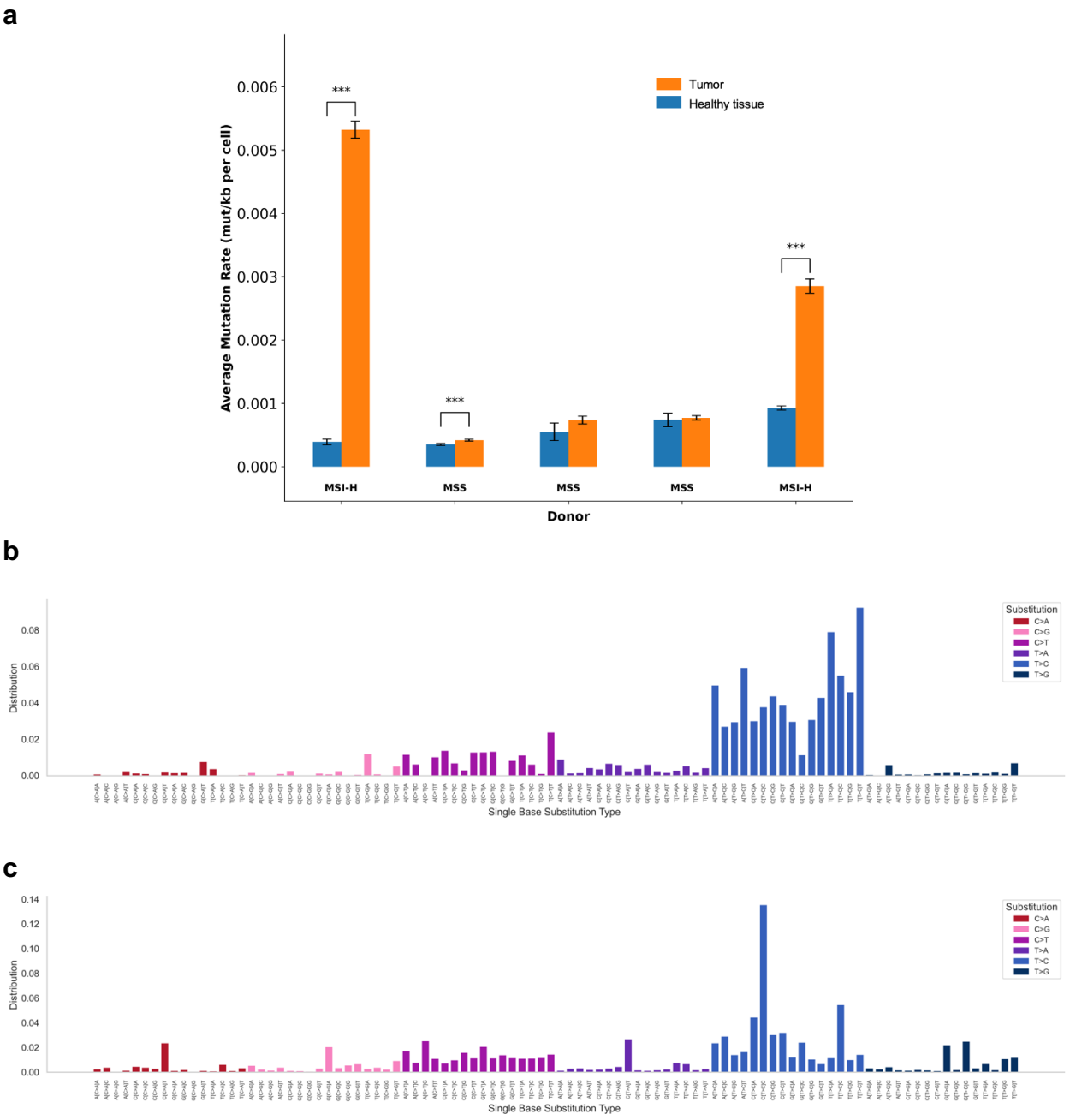
a



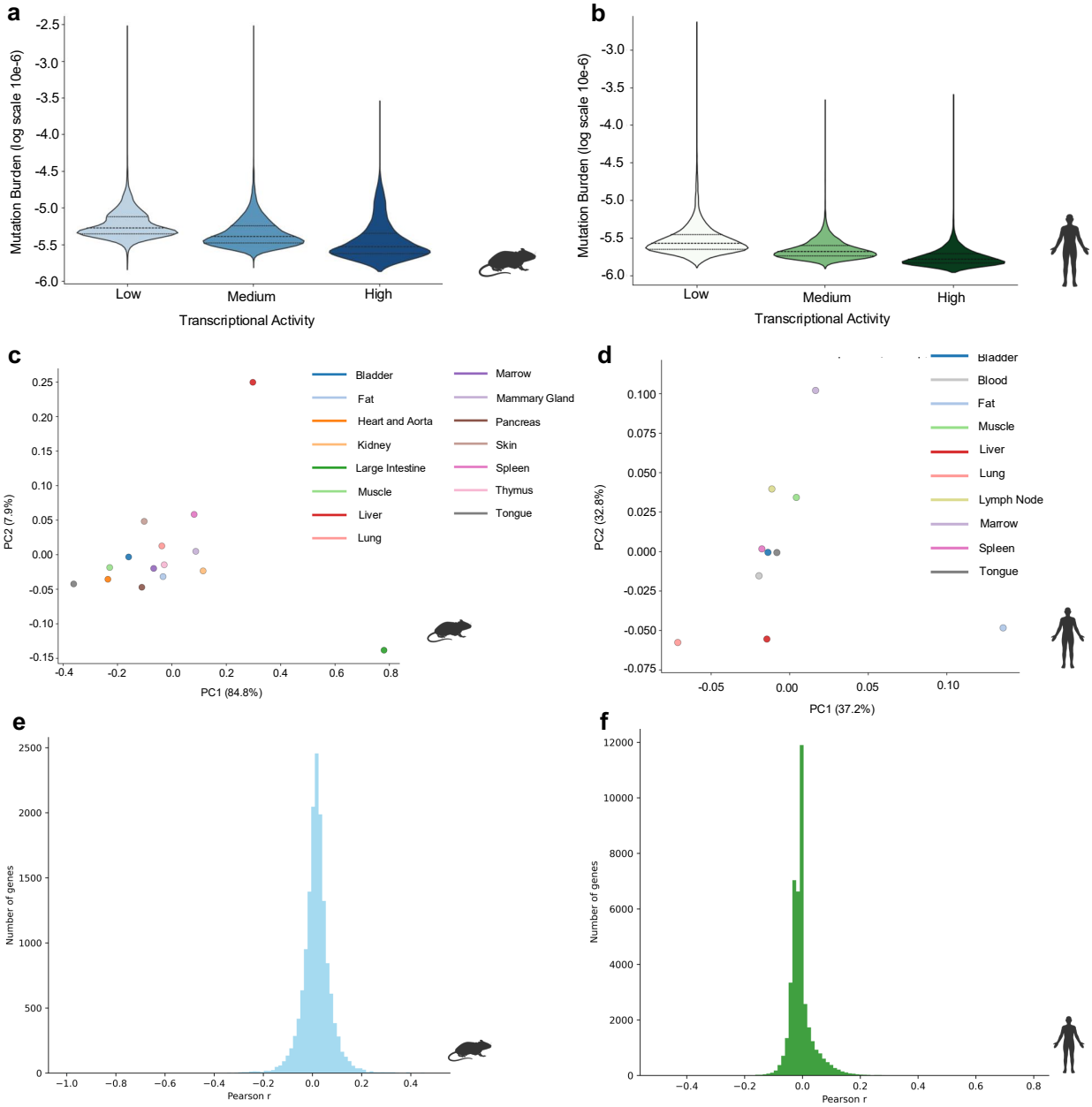
b



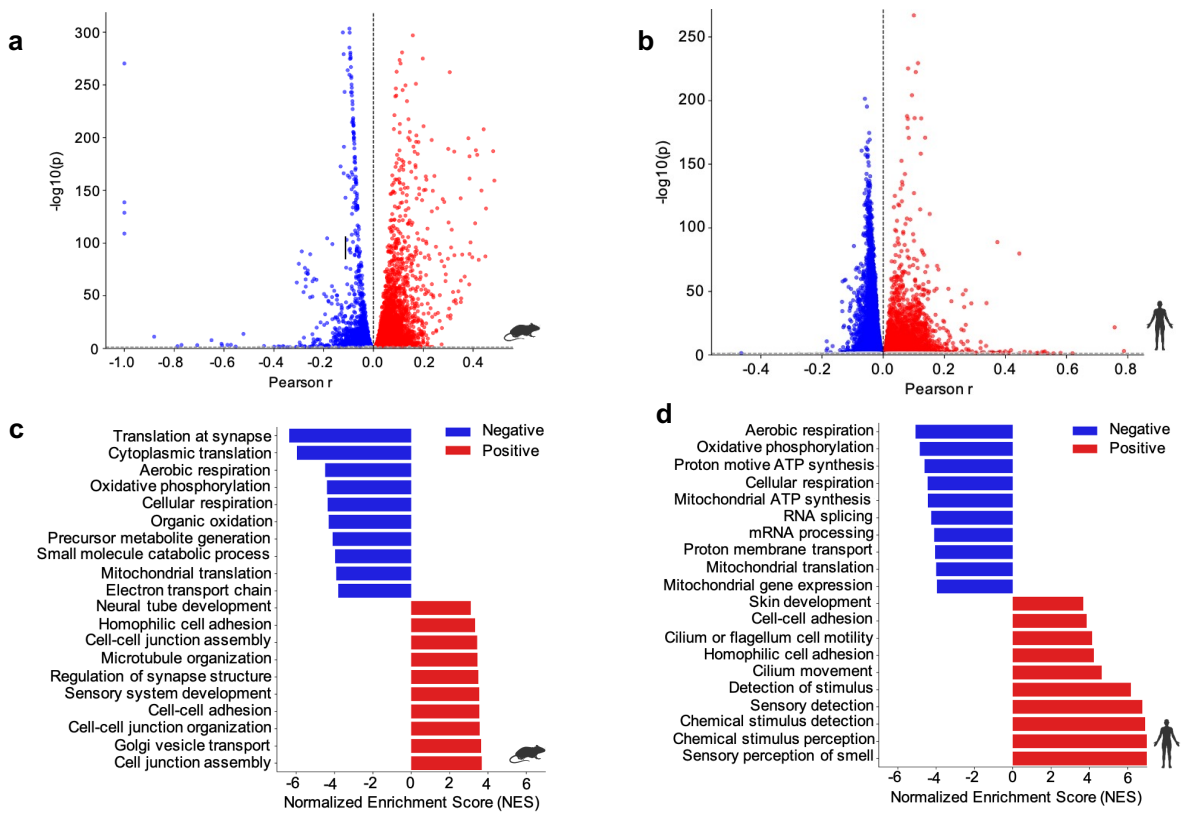
Extended Data Figure 2 | Developmental origin of somatic mutations. a, Boxplots showing mean contribution of mutations assigned to embryoblast, germ layer–specific, tissue-specific, and postnatal stages in mice. **b,** Boxplots showing mean contribution of mutations assigned to embryoblast, germ layer–specific, tissue-specific, and postnatal stages in humans.



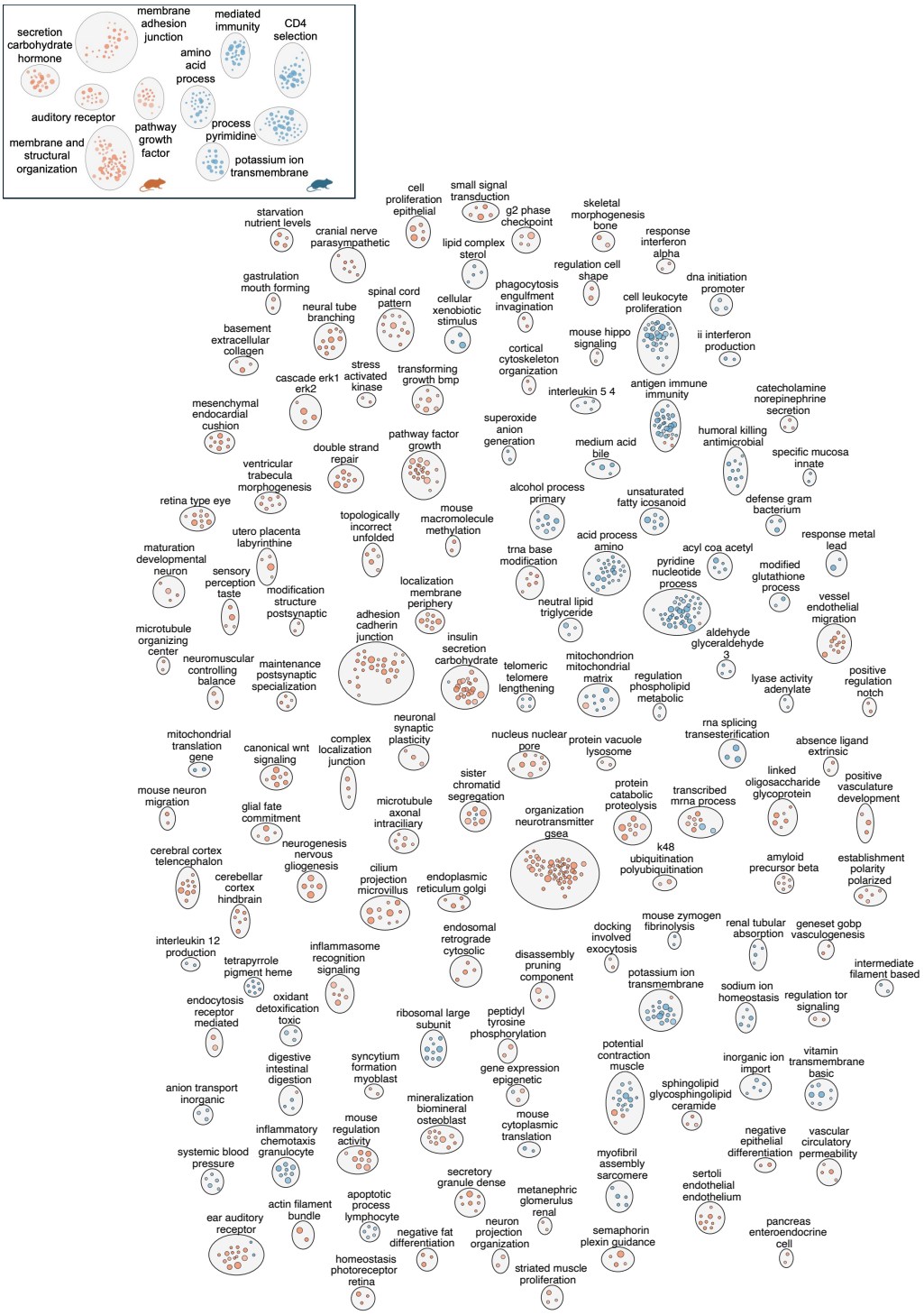
Extended Data Figure 3 | External Validation of the SComatic model. **a,** Average mutation rates (mutations per kilobase per cell) are shown for individual colon cancer donors stratified by microsatellite status (MSI-H vs. microsatellite stable, MSS). Healthy tissue (blue) and tumor tissue (orange bars) indicate per-donor samples, with error bars representing standard error of the mean (***p< 0.001). **b,** 96-class mutational spectra showing the distribution of substitutions across all trinucleotide contexts in mouse. **c,** 96-class mutational spectra showing the distribution of substitutions across all trinucleotide contexts in humans.



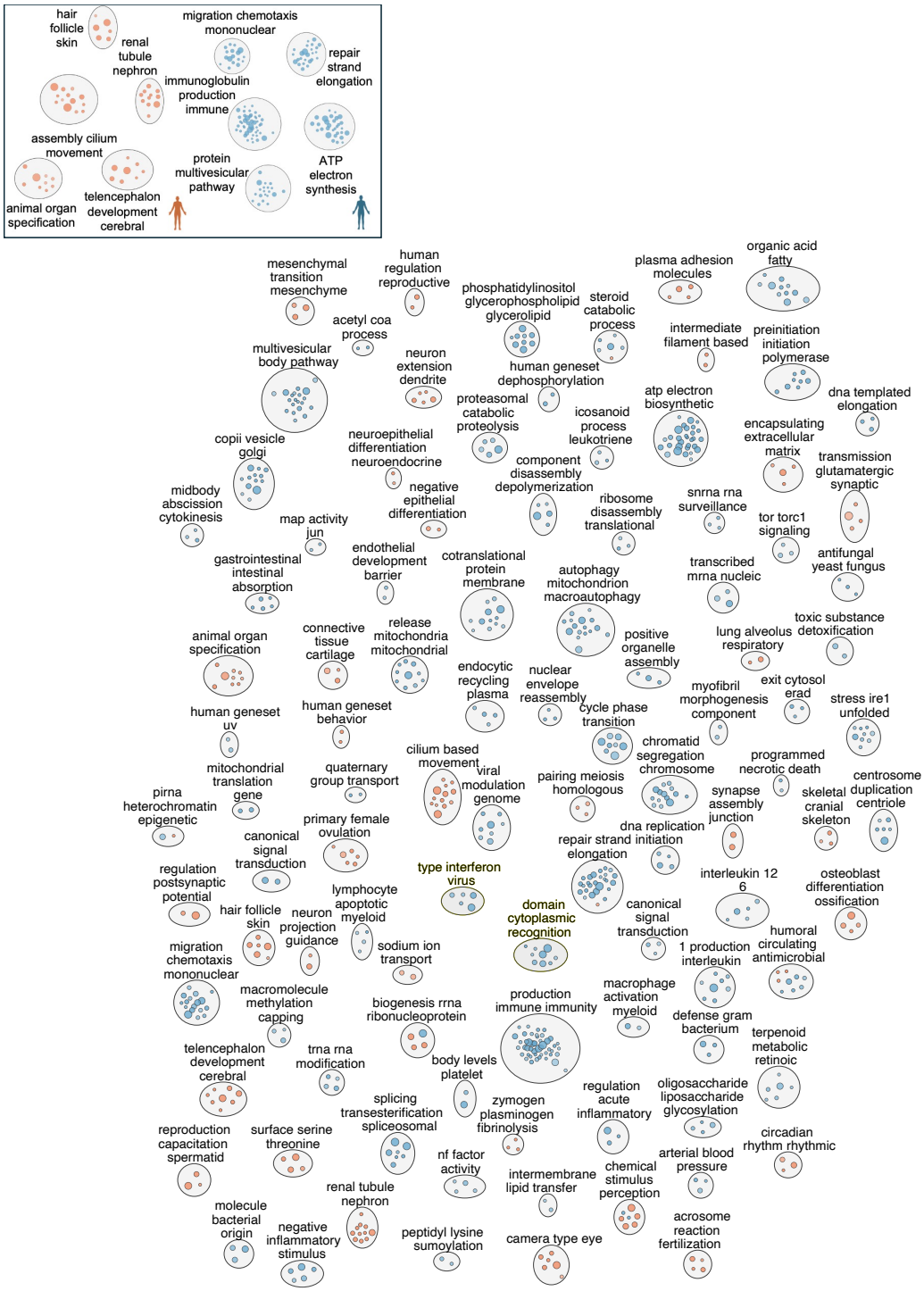
Extended Data Figure 4 | Transcriptional activity is associated with somatic mutation burden and genome-wide mutation–expression relationships across species. a, Violin plots showing per-cell mutation burden (log scale) stratified by transcriptional activity tertiles in mice. **b,** Violin plots showing per-cell mutation burden (log scale) stratified by transcriptional activity tertiles in humans. **c,** PCA of tissue-specific mutation spectra in mice. Each point represents the aggregate mutational profile of a tissue, with colors denoting tissue identity. **d,** PCA of tissue-specific mutation spectra in humans. Each point represents the aggregate mutational profile of a tissue, with colors denoting tissue identity. **e,** Histograms of Pearson correlation coefficients (r) between gene expression and mutation burden across all genes in mice. **f,** Histograms of Pearson correlation coefficients (r) between gene expression and mutation burden across all genes in mice.



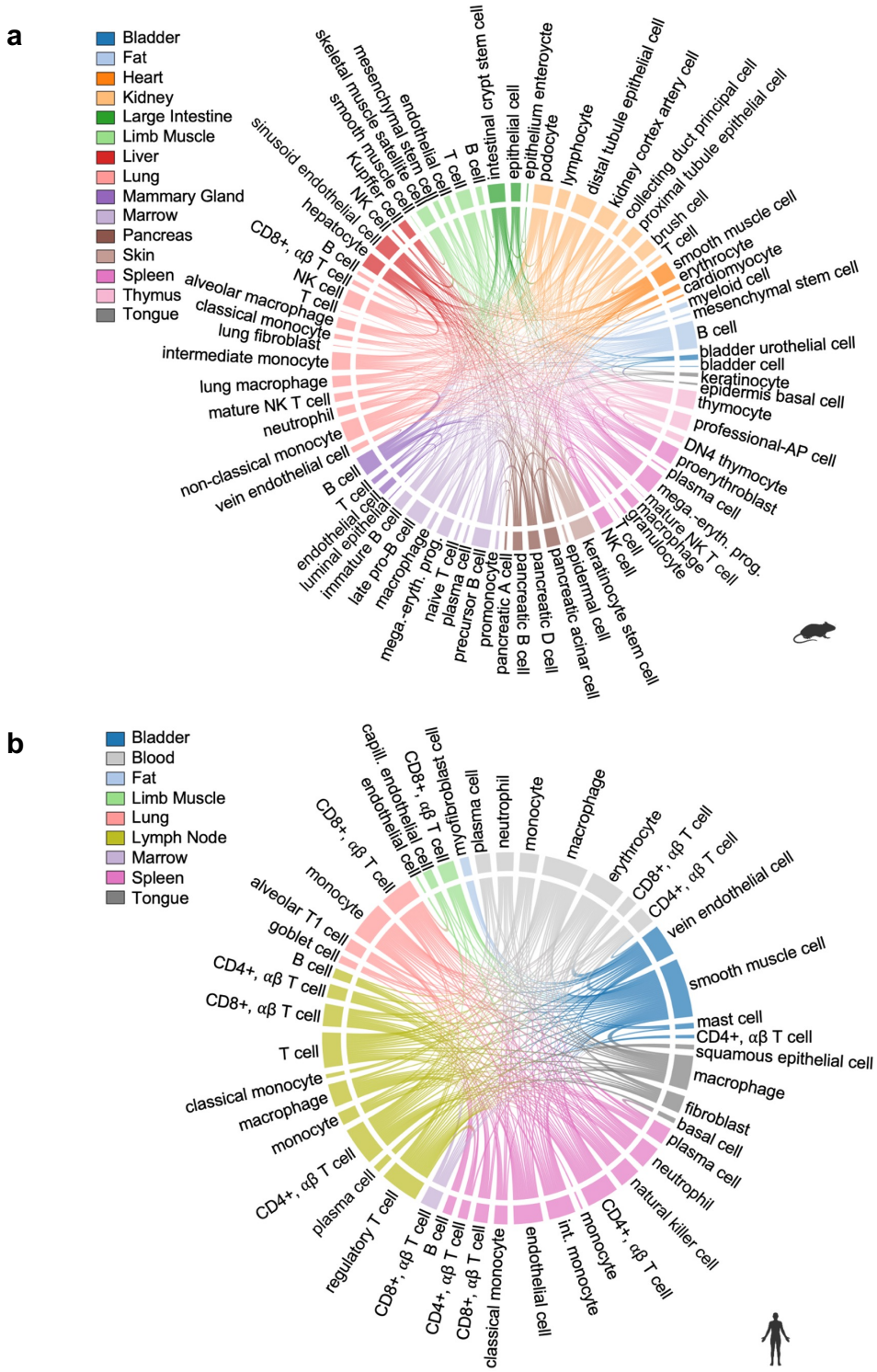
Extended Data Figure 5 | Gene expression–mutation correlations identify conserved functional pathways across species. a, Volcano plot showing the relationship between gene expression and somatic mutation burden in mice. Each point represents a gene; the x-axis indicates the Pearson correlation coefficient between expression and mutation burden, and the y-axis shows statistical significance (–log10 p-value). Genes with positive correlations are shown in red and those with negative correlations are shown in blue. **b,** Volcano plot in humans. **c,** GO enrichment analysis of expression–mutation correlations in mice. Bar plots display the top 10 positively (red) and negatively (blue) enriched pathways **d,** GO enrichment analysis of expression–mutation correlations in humans.



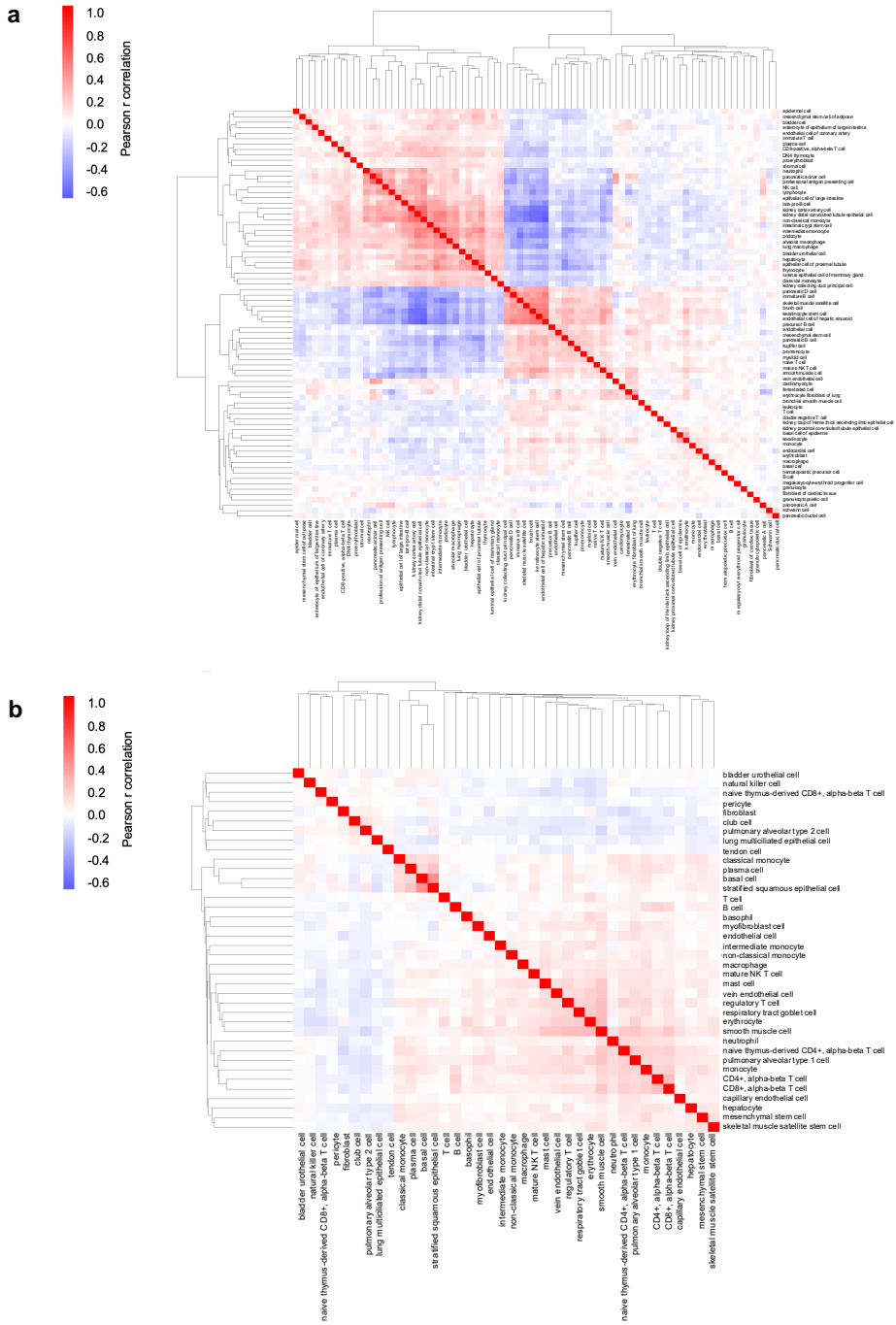
Extended Data Figure 6 | Global gene ontology enrichment network of transcriptional changes in mice. Enrichment map analysis of GO terms derived from gene expression associated with mutation burden of individual cells. Each node represents a GO term, with node size proportional to the number of associated genes and edges indicating functional similarity. Terms are colored by regulation status, with blue nodes denoting pathways downregulated with higher mutation burden and coral nodes denoting pathways upregulated with higher mutation burden. Top left boxed inset shows top 5 up- and top 5 down- regulated clusters.



Extended Data Figure 7 | Global gene ontology enrichment network of transcriptional changes in humans. Enrichment map analysis of GO terms derived from gene expression associated with mutation burden of individual cells. Each node represents a GO term, with node size proportional to the number of associated genes and edges indicating functional similarity. Terms are colored by regulation status, with blue nodes denoting pathways downregulated with higher mutation burden and coral nodes denoting pathways upregulated with higher mutation burden. Top left boxed inset shows top 5 up- and top 5 down- regulated clusters.



Extended Data Figure 8 | Lineage programs shape somatic mutation burden across tissues. **a,** Chord diagram shows commonly altered GO-terms associated with mutation burden across cell types in mice. Chord-width reflects larger overlap. **b**, Chord diagram shows commonly altered GO-terms associated with mutation burden across cell types in humans.



Extended Data Figure 9 | Cross-cell-type functional similarity based on shared GO-term enrichment. a, Pearson correlation heatmap showing functional similarity among mouse cell types, calculated from NES of significantly enriched GO terms. Each cell type's GO profile was derived from ranked gene expression values and summarized by its NES across biological processes. Correlations were computed between all pairs of cell types based on these NES vectors, reflecting the extent of shared functional programs. Red indicates positive correlation (shared enrichment for similar biological processes), and blue indicates negative correlation (divergent functional profiles). **b,** Pearson correlation heatmap showing functional similarity among human cell types, calculated from NES of significantly enriched GO terms.