

Supplementary Information

PDCD10 loss is associated with inflammatory and barrier-related endothelial remodeling in cerebral cavernous malformation

Cellular and Molecular Neurobiology

Xue Xiao^{2#}, Da Li^{3#}, Hengwei Song⁴, Yuxuan Ma⁴, Jiancong Weng^{1*}

Correspondence: Jiancong Weng, Department of Neurosurgery, China-Japan Friendship Hospital, Beijing 100029, China. jiancong.weng@sickkids.ca

Supplementary Methods and Results · Figures S1–S7 · Tables S1–S10

Supplementary Methods

Discovery dataset and experimental unit

Mouse single-cell RNA-sequencing count matrices were obtained from GSE155788. Four postnatal day 8 samples were analysed: P8wt1 (GSM4712077), P8wt2 (GSM4712078), P8CCM3KO2 (GSM4712079) and P8CCM3KO3 (GSM4712080). Wild-type and endothelial-specific *Pdcd10* knockout groups therefore each contained two biological replicates. The biological sample, not the individual cell, was treated as the replicate for genotype differential-expression analyses. Cell-level tests were used only for marker discovery, annotation or descriptive graph analysis.

Pre-processing, quality control and integration

Raw matrices were imported into Seurat. Genes detected in fewer than three cells were removed at object creation. Cells with fewer than 500 detected genes or at least 15% mitochondrial transcripts were excluded. Libraries were log-normalized to 10,000 counts per cell. Three thousand variable genes were selected, mitochondrial percentage was regressed during scaling and 50 principal components were calculated. Samples were integrated by reciprocal PCA. The first 30 integrated dimensions were used for neighbour-graph construction and UMAP. Initial Louvain clustering used resolution 0.5 and random seed 20260822. Quality-control distributions, doublet calls and retained-cell summaries are supplied in Supplementary Table S1.

Doublet detection and singlet re-analysis

Doublets were called independently within each sample using scDbtFinder on raw counts with sample-aware clustering and serial execution. Cells classified as singlets were merged, and normalization, variable-feature selection, scaling, PCA, reciprocal-PCA integration, neighbour construction and UMAP were repeated from the retained cells. Singlet clustering used resolution 0.5. Cluster markers were identified using two-sided Wilcoxon rank-sum tests with positive markers only, minimum detection fraction 0.20 and log2 fold-change threshold 0.25. To limit domination by large groups, at most 800 cells per identity were sampled for marker testing. Marker tests informed annotation and were not used as genotype-level inferential tests.

Broad annotation and endothelial selection

Broad identities were assigned from coordinated canonical markers for endothelial, immune, mural/perivascular, neural and erythroid compartments. Endothelial-enriched cells were extracted and re-analysed using 30 reciprocal-PCA dimensions and clustering resolution 0.7. Endothelial

cluster-marker testing used minimum detection fraction 0.15 and log2 fold-change threshold 0.20. Clusters dominated by erythroid, mixed or perivascular signals were excluded from the clean endothelial pseudobulk object. Annotation required coherence across multiple markers and programmes rather than expression of a single gene (Supplementary Fig. S1).

Vascular-zonation-first annotation

Arterial, arterial-capillary, capillary, venous-capillary and venous programmes were defined from the continuous brain vascular zonation framework of Vanlandewijck et al. Angiogenesis was scored separately as a functional programme because active angiogenesis can occur within more than one anatomical vascular segment. Pathological endothelial cells were not removed before origin inference (Supplementary Fig. S2).

Cells retaining endothelial identity and a validated five-segment zonation label served as anchors. For each non-anchor endothelial-lineage candidate, the 25 nearest anchors were identified in the first 30 reciprocal-PCA dimensions. Each neighbour contributed a weight of $1/(\text{distance}+10^{-5})$. The origin with the largest summed weight was assigned and confidence was calculated as the winning weight divided by total weight. Anchors retained their original label with confidence 1. This procedure estimates probable transcriptional origin; it is not lineage tracing. Confidence and ambiguous assignments are reported in Supplementary Table S2.

Endothelial pathological-state annotation

The endothelial–mesenchymal continuum was independently re-clustered at resolution 0.55 after normalization, mitochondrial regression, PCA integration and UMAP. Predefined modules captured endothelial identity, mesenchymal/ECM remodelling, proliferation, hypoxia, inflammatory/NF- κ B signalling, TGF- β /SMAD signalling, VEGF–ANGPT activity, leukocyte adhesion, coagulation/fibrinolysis and blood–brain barrier specialization. UCell was used for rank-based single-cell signature scoring (Supplementary Fig. S3; Supplementary Table S3).

Cells were assigned to endothelial identity retained, cycling endothelial, early EndMT, intermediate EndMT, cycling EndMT-like, identity-lost/fibroblast-like candidate, mural/pericyte or unresolved categories from combined module and marker evidence. Cycling EndMT-like required concurrent cell-cycle and EndMT evidence; proliferation genes alone were insufficient. The identity-lost/fibroblast-like label was deliberately conservative because no lineage tracing was available and the group was small. Stage proportions and module scores were summarized within each biological sample and treated descriptively.

Pseudobulk differential expression

Raw gene counts were summed by mouse across all retained clean endothelial cells and separately within endothelial subtypes. A subtype contrast was evaluated only when every sample contributed at least 30 cells. edgeR DGEList objects were filtered using filterByExpr with minimum count 10 and normalized by the trimmed mean of M values. A genotype-only design was fitted using robust dispersion estimation and quasi-likelihood testing. The tested coefficient was CCM3 knockout versus wild type. Benjamini–Hochberg FDR below 0.05 defined differential expression. Cell numbers are reported for transparency but do not increase the biological replicate count.

Hallmark gene-set enrichment

Mouse Hallmark genes were ranked by $\text{sign}(\log_2 \text{ fold change})$ multiplied by the square root of the edgeR quasi-likelihood F statistic. fgseaMultilevel used gene-set sizes of 15–500 and $\text{eps}=0$. Positive NES indicates enrichment in knockout. The all-endothelial contrast is shown in Supplementary Figure S4; the retained activated-mesenchymal contrast is shown in Figure 3b. This functional subtype contrast is not equivalent to the five-origin annotation. Complete retained tests are supplied in Supplementary Table S4.

UCell pathway summaries

Single-cell UCell scores were calculated from fixed signatures. For Figure 3, stored per-cell scores were joined by barcode to the six-state Figure 2 annotation, averaged within mouse and state, and then averaged within genotype. Heatmap rows were z-scaled only for display. Exact pathway gene memberships and sample-level summaries are supplied in Supplementary Table S3. Earlier functional signatures and summaries are separately identified. No cell-level genotype P values were used.

Monocle3 trajectory analysis

The combined 27,445-cell endothelial-lineage object was converted to a Monocle3 cell-data set while retaining the same UMAP coordinates used in the state map. WT and knockout cells were not fitted separately. Cells were clustered with $k=30$ and resolution 0.001. A principal graph was learned across partitions with minimum branch length 12. Eligible roots were WT cells annotated as Endothelial in the original state scheme. The root score was the standardized endothelial-identity score minus standardized early-EndMT and late-ECM scores. The top $\max(50, \text{ceiling}(0.05 \times \text{eligible cell count}))$ cells were selected. The retained starting set contained 404 cells mapping to 148 root principal nodes. This common root set was applied to both genotypes. Original root-cell flags and principal-node identifiers are provided in Supplementary Table S5. The graph was interpreted as a branched transcriptional organization (Supplementary Fig. S5).

For Figure 4b–g, finite-pseudotime cells from the retained shared graph were assigned to 30 approximately equal-cell bins. Expression was averaged first within mouse and bin and then by genotype. Mouse–bin summaries required at least 10 cells. The revised curves connect stored bin means without additional smoothing and combine cells across branches; no single disease branch was selected. Root-choice and fibroblast-like branch diagnostics are shown in Supplementary Figure S5. Additional retained branch/root correlations are provided in Supplementary Table S5. Pseudotime is not chronological time or lineage evidence.

Cell-cell communication inference

CellChat was applied separately to WT and knockout normalized expression matrices using the mouse ligand–receptor database. Included groups used the retained CellChat annotation: endothelial cells, cycling endothelial cells, early EndMT, intermediate EndMT, cycling EndMT-like, fibroblast-like candidates, mural/pericyte and broad immune populations. These grouped identities differ from the six-state atlas. Each sample-by-group combination was randomly capped at 500 cells, while rarer groups were retained in full. The final analysis included 9,964 cells. Overexpressed genes and interactions were identified, communication probability was calculated with the trimean estimator, no population-size weighting, 50 bootstraps and minimum group size 10, and pathway probabilities were aggregated. Differential values are knockout minus wild type. Model-derived

probabilities were not treated as mouse-level genotype tests (Supplementary Fig. S6; Supplementary Table S6).

Independent CCM3-loss validation

GSE246598 was restricted to the untreated, age-matched P10 comparison. Rapamycin-treated groups were excluded from the central validation analysis. The retained comparison comprised 3,352 *Pdcd10* knockout and 3,788 WT cells. Because each genotype was represented by a pooled library, gene effects were calculated from genotype-wise mean log-normalized expression and reported without replicate-level P values. Supporting comparisons are provided in Supplementary Table S7.

For GSE212018, the supplied normalized-expression table was transformed as $\log_2(\text{value}+1)$ and analysed as three WT versus three *Ccm3* knockout cerebellar endothelial samples using limma with empirical-Bayes moderation and trend estimation. Candidate genes and pathway modules were prespecified from GSE155788. Cross-dataset pathway summaries report mean and median gene-level $\log_2$ fold changes and are not a formal meta-analysis. Supporting comparisons are provided in Supplementary Table S7.

Cross-genotype organoid comparison

GSE276497 contains human induced-pluripotent-stem-cell-derived blood-vessel organoids after CRISPR inactivation of *CCM1*, *CCM2* or *CCM3*. The analysis used the supplied integrated object and pairwise differential-expression table for endothelial cluster c9. Differential genes were counted using the common supplied threshold of absolute normalized average log fold change >0.5 and adjusted $P < 0.05$. One pooled library was available for each genotype; results were therefore treated descriptively. DEG counts describe the supplied comparisons and are sensitive to cell numbers, detection power and thresholds. They were not used as replicated tests of genotype effects or as measures of lesion or hemorrhage severity. Original c9 comparison tables and displayed effects are provided in Supplementary Table S8.

Statistical analysis and limitations

All replicate-level genotype models used the biological sample as the experimental unit. Multiple testing was controlled with the Benjamini–Hochberg method. Composition, UCell scores, pseudotime distributions, pooled-library contrasts and CellChat outputs were descriptive. Vascular origins for identity-lost cells were probabilistic. EndMT was inferred from transcriptional state rather than lineage tracing. Trajectory analysis could not establish chronological progression, and communication analysis could not establish physical signalling. Cross-dataset agreement strengthened reproducibility but did not remove differences in developmental stage, tissue processing, species or platform.

Software and reproducibility

Recorded analysis sessions used R 4.5.2, Seurat 5.5.1, SeuratObject 5.4.0 and Monocle3 1.4.27. Per-step session records document the remaining package versions. The revision scripts, source-data manifest and session records are available from the corresponding author on request. Original analyses and the figure-summary revisions are distinguished in that manifest.

Relationship between annotation systems

Main Figures 1–4 and Supplementary Figures S2–S3 use the harmonized origin and six-state annotations. Supplementary Table S2 provides current cell assignments and explicitly marked earlier cluster annotations. Supplementary Table S4 retains earlier functional-subtype contrasts, including an angiogenic subtype; these are not five-origin contrasts. Supplementary Figure S5 preserves the labels of the earlier root-sensitivity analysis. CellChat retains its original eight groups, including separate endothelial and cycling-endothelial groups. These population schemes must not be treated as numerically interchangeable. The original discovery models and Figures 1–7 were retained. The additional external count models are described separately below.

Clinical relevance and experimental context

Evidence was organized by proximity to human disease, not by a validated numerical hierarchy. Patient-lesion tissue would provide direct clinical context; no genotype-resolved human hemorrhage comparison was available in the analyzed set. Mouse brain endothelial comparisons formed the principal model-based evidence. Human vascular organoids supplied exploratory multicellular and cross-genotype context, while cultured endothelial and guide-perturbation comparisons tested responses in more restricted settings. Patient-derived cultured cells were distinguished from uncultured lesion specimens. Within each context, biological replication, control matching, tissue and genotype annotation, and data provenance were evaluated separately. An unreplicated human-derived model was not treated as stronger inferential evidence simply because of its species.

GEO was the source of the completed quantitative comparisons, but ArrayExpress/BioStudies and ENA were also searched. Repository membership and agreement with the discovery result did not determine eligibility. The public E-MTAB-13726 record supplied raw reads requiring reconstruction; reconstruction was incomplete, without a documented permission restriction. GSE186523 supplied expression files but had conflicting title and condition annotations. Both records remain visible in Table S10. These limitations reduce coverage and do not establish absence of a biological effect.

Retrieval and assessment summary

Searches conducted on 9 September 2026 retrieved 346 GEO series, 44 ArrayExpress accessions and 63 ENA study records after within-source accession deduplication. The following accounting summarizes records rather than independent cohorts; linked deposits and superseries are not additional biological replication.

GEO: 346 records → 304 excluded at initial screening + 4 duplicate or superseries records + 38 full assessments.

ArrayExpress and ENA: 107 records → 81 linked records + 25 exclusions + 1 additional full assessment (E-MTAB-13726).

Full assessment: 39 records → 18 eligible + 9 descriptively eligible + 3 contextual + 8 excluded + 1 unresolved sample-label record (GSE186523).

Eligibility is distinct from completed analysis. The 18 eligible records include E-MTAB-13726, for which public raw-read reconstruction was incomplete at the reporting cutoff. No quantitative effect from that record was included. The register retains both this eligible but unanalyzed record and the unresolved GSE186523 comparison.

Dataset search and eligibility

GEO, the ArrayExpress collection within BioStudies, and ENA were searched on 9 September 2026. Disease and gene-name queries, returned records, repository identifiers and item-level decisions are supplied in Supplementary Table S10. The screening scope was human or mouse endothelial gene-expression comparisons involving KRIT1/CCM1, CCM2 or PDCD10/CCM3 loss with an untreated or vehicle-matched control. Gene-expression platforms and endothelial culture, isolated-cell and organoid models were considered. Non-endothelial tissue, non-expression assays, other organisms and experiments without a relevant loss-control contrast were excluded from the central synthesis. Genotype-unresolved human disease observations were recorded separately as contextual evidence. Eligibility was not selected on direction, significance or agreement with the discovery findings.

All 346 GEO retrieval records received title/design screening. Thirty-eight proceeded to sample-level assessment; four superseries were retained as linked records rather than independent experiments. The 44 ArrayExpress and 63 ENA records were mapped to canonical accessions where explicit identifiers established overlap. E-MTAB-13726 contributed a primary non-GEO record. Similar study titles alone were not taken as proof of independence or duplication. The full assessment register contains 39 records: 27 eligible or descriptively eligible, three contextual, eight excluded and one with unresolved sample labels. These are record counts, not counts of independent patient or animal cohorts. Non-GEO raw reconstruction remains outstanding and is explicitly identified in Table S10; the quantitative expansion is not an exhaustive meta-analysis of all eligible experiments.

Sample selection

GSE123968: primary brain microvascular endothelial cells from adult mice, analyzed as separate Krit1-loss and Pdc10-loss comparisons, each with three loss and three matched-control mice. Genotype was assigned from GSM3517832–GSM3517843 sample titles and genotype fields. Cultured cells were harvested at different intervals for the two genes, and sequencing read lengths differed. These contrasts are not a controlled test of clinical genotype severity (Koskimäki et al. 2019; doi:10.1172/jci.insight.126167).

GSE204979: the seven brain endothelial libraries MBECF2–4 and MBECK1–4 were retained, corresponding to GSM6203353–GSM6203359. Astrocyte libraries and RPKM columns were not included in the raw-count model. The comparison comprised four Pdc10-loss and three floxed-control mice, with age P75 in the deposited metadata (Lai et al. 2022; doi:10.1161/CIRCRESAHA.122.321129).

GSE261224: only the nine Normoxia columns were retained, comprising four Flox controls and five KO samples, GSM8138083–GSM8138091. Hypoxic samples were not pooled into this contrast (Frias-Anaya et al. 2024; doi:10.1161/ATVBAHA.123.320367).

Count models and enrichment

Deposited gene counts were checked for nonnegative integer values, finite entries and unique Ensembl identifiers. Ensembl version suffixes were removed where present. edgeR filterByExpr retained sufficiently expressed genes within each comparison. TMM normalization, robust dispersion estimation and robust quasi-likelihood models tested loss versus control. Benjamini–Hochberg FDR was calculated over all genes tested in each contrast, not only the displayed candidates. No sample was dropped based on its differential-expression result.

Mouse Ensembl identifiers were mapped to symbols with org.Mm.eg.db; the complete mapping is retained with the analysis. When multiple Ensembl genes mapped to a symbol, the highest mean-abundance gene was used for enrichment, not the gene with the largest differential statistic. fgseaMultilevel used signed square-root F statistics, the existing mouse Hallmark collection, set sizes 15–500, eps=0, and seed 20260909. All tested pathways, gene-level outputs and sample columns are included in the Supplementary Table S9. Original-count models, package versions and scripts are retained with the analysis.

For GSE123968, legacy edgeR quasi-likelihood and robust limma-voom analyses assessed model sensitivity using all six samples per genotype-specific comparison. The absence of FDR-significant genes in the cultured Pdcd10 comparison persisted under both alternative methods. This is a result under the stated models, not proof of biological equivalence or absence of gene-loss effects.

Human endothelial culture comparisons

GSE165406: GSM5032901–GSM5032903 were non-targeting-siRNA controls and GSM5032904–GSM5032906 were CCM2-only knockdown samples. ROCK co-silencing samples were not pooled. Deposited headerless gene-symbol/count files were joined after confirming identical gene order and unique symbols. Three culture replicates per group were modeled without an invented pairing factor.

GSE329903: vehicle-treated shScramble GSM9712927–GSM9712929 and shCCM3 GSM9712921–GSM9712923 were retained. The deposited batch values 1–3 were included in the design (~batch + condition). Dabrafenib-treated samples were not included. Gene-level featureCounts were mapped with org.Hs.eg.db. Both human comparisons used filterByExpr, TMM and robust edgeR quasi-likelihood tests with BH correction across all tested genes. Donor independence is not established by culture-replicate labels. Human and mouse Hallmark analyses are now supplied with species-specific signatures in Table S9.

CCM2 knockdown increased SERPINE1, PGF and NFKBIA and reduced MFSD2A. The CCM3 vehicle contrast increased PGF and NFKBIA and reduced MFSD2A, but decreased SERPINE1 and FN1. These differences preclude a universal CCM3-specific effector signature. Complete human outputs are included in the Supplementary Table S9; Fig. S7 displays the four mouse contrasts only.

Interpretation and unresolved data issues

Two additional brain endothelial comparisons supported increased Serpine1, Pgf and Fn1 and positive enrichment of inflammatory, hypoxic, TGF- β and coagulation programmes. Significant reductions in both barrier candidates were not reproduced. The cultured Pdcd10 comparison did not reproduce the candidate expression pattern. Both concordance and nonconcordance are retained in the manuscript.

GSE186523 currently has a conflict between sample titles/file names and condition annotations. Aortic endothelial silencing and three culture replicates per condition are described in its source paper, but this does not independently resolve the accession-to-condition mapping. Directional interpretation remains withheld pending clarification; labels are not reversed to match an expected result (Fusco et al. 2022; doi:10.3390/genes13060961).

E-MTAB-13726 contains a three-versus-three DMSO-matched WT/KO contrast and links to ENA ERP156795/PRJEB72007. The repository record has raw-read access but no processed expression matrix. Raw-read reconstruction is not complete. This eligible record was not included in the quantitative synthesis; the resulting coverage limitation is retained in the assessment register.

Additional clone and deposited-result comparisons

GSE137425 supplied gene-level integer counts for WT1, WT2 and WT4 and CCM3-null clones K17, K18 and K19. Counts were checked for finite nonnegative integer values and unique genes, then analyzed with filterByExpr, TMM, robust edgeR quasi-likelihood estimation and a loss-versus-control coefficient. Three lines or clones per group do not establish independent human donors. Complete tests, candidate effects and sample columns are supplied in Supplementary Table S9. The full deposited differential-expression tables for endothelial-stage GSE214306, DMSO-matched GSE310745 and GSE85657 were extracted separately. Their supplied effect directions were checked against group-labelled expression columns. Author-reported adjusted P values were retained as reported, not described as independently recomputed tests. Fractional normalized expression was not rounded into raw counts. GSE85657 transcript-level entries were retained without choosing the most significant transcript per gene.

Further count and normalized-expression models

Untreated WT1–WT3 and CCM3-null K21–K23 libraries from GSE137426 were analyzed separately from GSE137425. They are different deposited libraries from the same source study, not an independent donor cohort. Four GSE112882 HUVEC hairpin contrasts (CCM1-2, CCM2-1, CCM2-5 and CCM3-2) used the two matched control cultures and excluded Y27632 treatment. Deposited gene identifiers converted by Excel into dates were excluded with an explicit audit rather than assigned guessed gene symbols. GSE298723 comprised three untreated WT and three untreated knockout libraries. Salmon transcript identifiers matched GENCODE mouse M27 annotations; tximport lengthScaledTPM gene estimates were analyzed with robust edgeR quasi-likelihood models without rounding them into integer read counts. These additions bring the expanded count or estimated-count analysis to 13 contrasts.

The chronic GSE246373 cohort included five knockout and five WT samples. Its acute arm reused GSE212018 and was not counted twice. Deposited normalized counts were transformed as $\log_2(\text{value}+1)$, filtered at expression of at least 10 in five samples and analyzed with robust limma-trend. GSE54095 normalized microarrays were analyzed separately at passages 7 (four samples per group) and 11 (five per group), without an additional logarithmic transformation. Probes were filtered at expression of at least 6 in the minimum group size; the deposited GPL6244 annotation was retained, without selecting the most significant probe for a gene. These three models are supplied separately from count tests in Table S9.

Descriptive and author-reported evidence

GSE210522 and GSE232400 guide comparisons used matched conditions and mean $\log_2(\text{TPM}+1)$ differences. GSE291768 mutant-line comparisons against the available healthy-control line used mean $\log_2(\text{TMM CPM}+1)$; culture repeats were not treated as independent donors. GSE288811 used all deposited libraries within matched flow conditions and mean $\log_2(\text{normalized expression}+1)$. GSE112882 HCMEC contexts used $\log_2(\text{RPKM}+1)$. No new biological-replicate P values were assigned to these descriptive comparisons. Missing or unavailable genes were recorded rather than replaced with zero.

All nine deposited GSE116323 GFOLD comparisons were retained by gene and matched culture/time context. GFOLD E-FDR was not relabelled as Benjamini–Hochberg FDR. GSE212396 and GSE210681 supplied CCM2 and KRIT1 guide effects; their author-reported cell-level P values do not establish donor-level inference. PDCD10 was absent from the supplied GSE210681 effect

matrix. For GSE246174, 509 CCM2, 461 WT-in-CCM2-mosaic and 464 WT-in-control-mosaic cells were matched to the deposited post-QC count matrix and summarized by mean log₁₀ expression normalized to 10,000 counts and fraction detected. The three pooled libraries differ in mosaic context and sorting, so no independent genotype significance test was performed.

Expanded Hallmark enrichment

Hallmark analysis was performed for the 13 count models and three normalized-expression models. Species-matched signatures and their versioned memberships are retained with the analysis. Count-model genes were ranked by signed square-root quasi-likelihood F statistics and normalized-expression models by moderated t statistics. Where multiple features mapped to one symbol, the feature with the greatest average expression was retained, not the greatest test statistic. fgseaMultilevel used gene-set sizes 15–500, $\text{eps}=0$ and a fixed seed. Adjusted P values refer to Hallmark tests within each model. Different models were not pooled into a cross-genotype severity score.

Supplementary Results

Reproducibility across endothelial models

The main synthesis emphasizes brain endothelial comparisons because they match the discovery tissue. The following detailed results retain concordant, nonsignificant and opposite-direction observations. They are presented by experimental context and were not used to determine dataset eligibility. Complete results are available in Supplementary Tables S8 and S9, with the full eligibility register in Table S10.

In GSE204979, *Serpine1*, *Pgf* and *Fn1* increased by log₂ fold changes of 4.13, 1.46 and 1.45, respectively (FDR=0.00090, 0.0050 and 0.00042). The normoxic GSE261224 comparison also showed increases in these genes (log₂ fold changes 2.57, 2.25 and 1.30; FDR=0.033, 0.0044 and 0.00047). Both comparisons showed positive enrichment of TNF–NF- κ B, hypoxia, TGF- β and coagulation Hallmark programmes. Neither comparison established significant reductions in *Cldn5* or *Mfsd2a* at FDR<0.05 (Supplementary Fig. S7; Supplementary Table S9).

By contrast, the cultured *Pdcd10*-loss comparison in GSE123968 did not reproduce significant changes in the six displayed candidate genes. No individual gene passed FDR<0.05 under the primary robust quasi-likelihood model, and legacy edgeR and limma-voom sensitivity analyses gave the same absence of FDR-significant genes. This is not evidence that *Pdcd10* loss has no biological effect. Coagulation enrichment remained detectable (NES=1.64; FDR=0.020), whereas TNF–NF- κ B and hypoxia were not significantly enriched. Thus, reproducibility was selective across experimental contexts; the comparisons do not establish a universally stronger CCM3 programme.

Human endothelial cultures further demonstrated model dependence. In GSE165406, CCM2 knockdown increased *SERPINE1*, *PGF* and *NFKBIA* (log₂ fold changes 1.48, 1.67 and 1.00; FDR=0.00023, 0.000059 and 0.00037), while *MFSD2A* decreased. In vehicle-matched GSE329903, CCM3 knockdown increased *PGF* and *NFKBIA* and reduced *MFSD2A*, but *SERPINE1* and *FN1* decreased (log₂ fold changes –1.23 and –0.98; FDR=1.6 $\times 10^{-7}$ and 2.4 $\times 10^{-7}$). Thus, selected responses were shared with CCM2 loss, whereas other candidates changed in opposite directions across models (Supplementary Table S9). In the additional GSE137425 clone comparison, *FN1* decreased (log₂ fold change –2.40; FDR=0.000027), while *NFKBIA* increased and *MFSD2A*

decreased. SERPINE1 and PGF did not meet $FDR < 0.05$, further limiting generalization across endothelial models. The chronic GSE246373 mouse cohort also supported increases in Serpine1, Pgf and Fn1 (log2 expression differences 2.68, 1.62 and 1.07; adjusted $P = 0.0020$, 0.0025 and 0.0062). Conversely, untreated GSE298723 did not show FDR-significant changes in the available displayed candidates, and GSE137426 and the GSE112882 CCM3 hairpin comparison showed decreases in SERPINE1 and PGF. These contrasting results were retained rather than used as exclusion criteria (Supplementary Table S9).

In the supplied human blood-vessel organoid endothelial cluster c9, the common reporting threshold identified 41 differentially expressed genes after CCM1 loss, 96 after CCM2 loss and 193 after CCM3 loss (Fig. 7a). The candidate-gene comparison showed shared changes in KLF2, KLF4 and THBD, with relatively prominent FN1 induction and CLDN5 reduction after CCM3 loss (Fig. 7b). Each genotype was represented by one pooled library, so the DEG counts describe this dataset and are not a biological-replicate test of differences in response breadth (Supplementary Table S8).

The organoids did not reproduce all mouse candidate signals: PGF decreased in CCM3-deficient organoids, SERPINE1 showed no comparable induction, and MFSD2A was unavailable in the selected comparison. Thus, the cross-genotype dataset supports shared CCM-related responses with model-dependent differences. It does not demonstrate that CCM3 uniformly strengthens the same inflammatory, fibrinolytic or barrier programme, nor does the absence of a detected exclusive pathway exclude CCM3-specific biology.

Supplementary Figure Legends

Supplementary Figure S1 | Quality control and analysis subsets

a, Raw, post-quality-control and singlet cell counts for each mouse. b,c, Detected genes and mitochondrial-transcript fractions in retained singlets. Boxes indicate the interquartile range, centre lines the median and whiskers observations within 1.5 interquartile ranges; outliers are omitted from display only. d, Stored singlet-atlas UMAP with the earlier broad endothelial label distinguished from other broad classes. e, Numbers of stringent endothelial cells used for pseudobulk and broader endothelial-lineage candidates used for origin/state analyses. These are analysis subsets, not sequential filtering steps. WT1, WT2, KO2 and KO3 identify the four biological samples; $n = 2$ mice per genotype. Sources: Supplementary Tables S1 and S2.

Supplementary Figure S2 | Marker evidence and confidence for inferred vascular origin

a, Vascular-zonation marker profiles across the five current vascular-origin assignments in the 27,445-cell lineage-candidate atlas. Colour indicates the gene-wise z score of mean log-normalized expression, clipped at -2 and $+2$; dot area increases with the fraction of cells with detected expression. b, Empirical cumulative distributions of origin-vote confidence. The dotted line marks 0.70 for reference, not a filtering step in this display. Anchors were assigned confidence 1 by construction; this is not calibrated lineage certainty. Origin labels represent transcriptomic affiliation. Source: Supplementary Table S2.

Supplementary Figure S3 | Extended state markers and mouse-level composition

a,b, Endothelial/matrix-associated markers (a) and proliferation/mural markers (b) across the same six states used in main Figure 2. Colours indicate gene-wise z scores of mean log-normalized expression, clipped at -2 and $+2$; dot area increases with the detected fraction. c, State fractions among all endothelial-lineage candidates within each mouse. The four bars are biological samples, not pooled genotype replicates. Cycling is a concurrent programme, not a later chronological stage. Fibroblast-like candidates have unresolved ancestry. Data and denominators are provided in Supplementary Table S2; programme definitions are in Supplementary Table S3.

Supplementary Figure S4 | All-endothelial differential expression and pathway enrichment

a, All-endothelial knockout-versus-WT pseudobulk comparison. The x axis shows $\log_2$ fold change and the y axis $-\log_{10}$ Benjamini–Hochberg FDR. Significant increased and decreased genes are red and blue, respectively; other genes are grey. The dotted line denotes $FDR=0.05$. Selected candidate genes are labelled. b, Hallmark pathways with $FDR<0.05$ in this all-endothelial contrast, ordered by normalized enrichment score (NES). Positive NES indicates knockout enrichment. Each genotype has two biological mouse replicates. This contrast differs from the activated-mesenchymal contrast in main Figure 3b. All retained tested genes and pathway results, including nonsignificant results and separately labelled subtype analyses, are provided in Supplementary Table S4.

Supplementary Figure S5 | Root sensitivity and the position of fibroblast-like candidates

a, State-level median scaled pseudotime under the retained multi-root solution and two alternative roots (RootA and RootB). Labels summarize the original root-sensitivity categories, which are not identical to the harmonized six-state atlas; their exact names are retained in Supplementary Table S5. b, Original multi-root pseudotime of fibroblast-like candidates on root-proximal and distal branches, distinguished by genotype. Each point is one candidate cell; horizontal jitter is for visibility. Their presence in both proximal and distal positions argues against assigning a uniform terminal stage. Pseudotime is branched transcriptional ordering, not elapsed time or lineage tracing. Root cells, root principal nodes, graph membership, sensitivity summaries and mouse-bin gene expression are supplied in Supplementary Table S5.

Supplementary Figure S6 | Global context for predicted cell-cell communication

a, The 12 pathways with the highest combined WT-plus-knockout information flow in the retained CellChat analysis, shown separately by genotype. Information flow is summed predicted communication probability. b, Sender–receiver differences obtained by summing knockout-minus-WT ligand–receptor probability differences over the retained interaction table. Red and blue indicate larger predicted values in knockout and WT, respectively. The original eight CellChat groups are retained, including separate endothelial and cycling-endothelial groups, pathological groups, mural/pericyte cells and immune cells. Grouping differs from the harmonized six-state atlas. These are model-derived summaries, not mouse-level significance tests or measurements of physical signalling. Complete outputs and the displayed matrix are provided in Supplementary Table S6.

Supplementary Figure S7 | Candidate effects across additional mouse endothelial datasets

Fig. S7 Candidate gene effects in additional endothelial datasets. Panels show *Serpine1*, *Pgf*, *Fn1*, *Nfkb1a*, *Cldn5* and *Mfsd2a* $\log_2$ fold changes for loss versus matched controls. Filled circles indicate gene-level BH $FDR<0.05$; open circles indicate $FDR\geq 0.05$. Numbers on the y-axis give loss-versus-

control sample counts. GSE123968 CCM1 and CCM3 cultured BMEC comparisons are separate models with different harvest intervals. GSE204979 and GSE261224 represent additional brain endothelial contexts; only normoxic GSE261224 samples were analyzed here. All panels share the same effect-size axis. No pooled effect, confidence interval or genotype-severity ranking is implied. Complete results and source data are provided in Supplementary Table S9.

Supplementary Table Index

Supplementary Table S1 | Datasets, QC and retained-cell counts

Supports: Methods; Figure 1; Figure S1.

Supplementary Table S2 | Cell annotations, marker evidence and composition

Supports: Figures 1–2; Figures S1–S3.

Supplementary Table S3 | Programme definitions and mouse-level scores

Supports: Figures 2–3; Figure S3.

Supplementary Table S4 | Complete retained differential-expression and enrichment results

Supports: Figure 3; Figure S4.

Supplementary Table S5 | Trajectory, original roots and sensitivity analyses

Supports: Figure 4; Figure S5.

Supplementary Table S6 | Predicted communication and interaction outputs

Supports: Figure 5; Figure S6.

Supplementary Table S7 | Independent mouse-dataset validation

Supports: Figure 6.

Supplementary Table S8 | Human organoid cross-genotype comparisons

Supports: Figure 7.

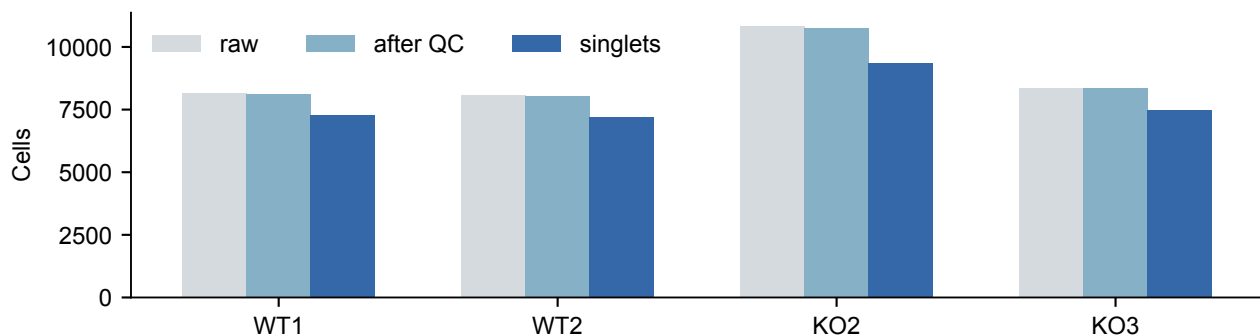
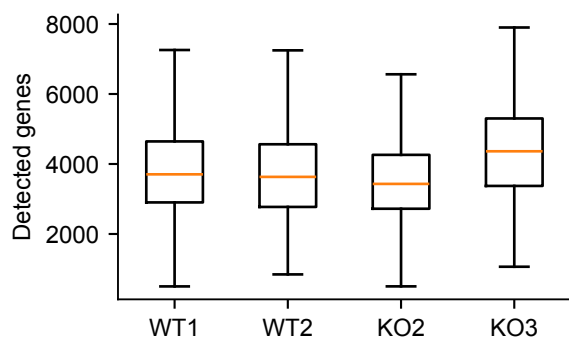
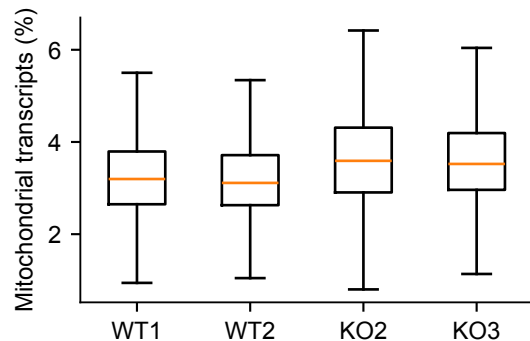
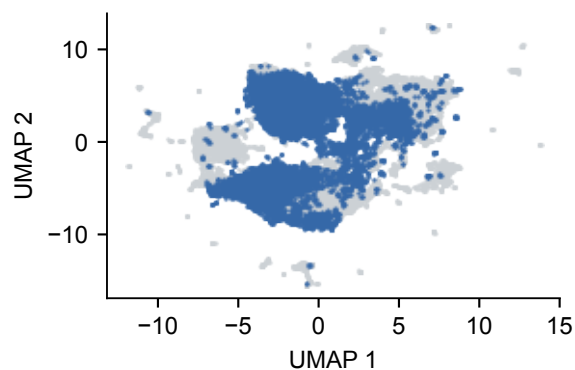
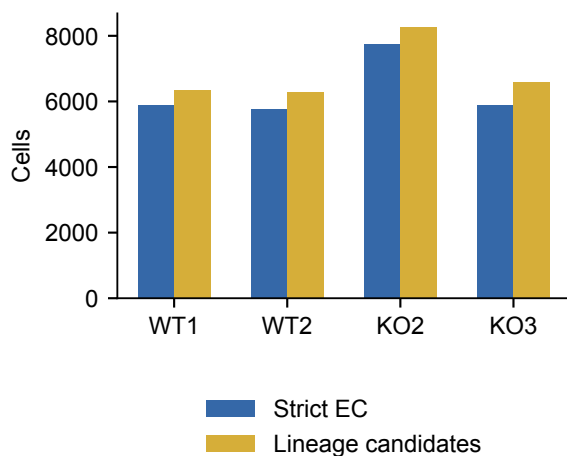
Supplementary Table S9 | Expanded external comparisons

Supports: Additional endothelial comparisons; Independent brain endothelial datasets support selected candidate responses; Supplementary Results; Supplementary Fig. S7. The workbook contains candidate effects, pathways, samples and sensitivity checks. Accompanying CSV parts separately contain complete count-model tests, normalized-expression tests, descriptive effects, author-deposited differential results, unreplicated GFOLD results, pilot and full guide-screen effects, and pooled expression. These are components of Table S9, not separate numbered tables.

Supplementary Table S10 | Dataset retrieval and screening

Supports: Dataset search and eligibility. Includes all GEO retrieval decisions, sample-level eligibility assessments, cross-repository links and search queries. Outstanding analyses and unresolved labels are explicitly retained.

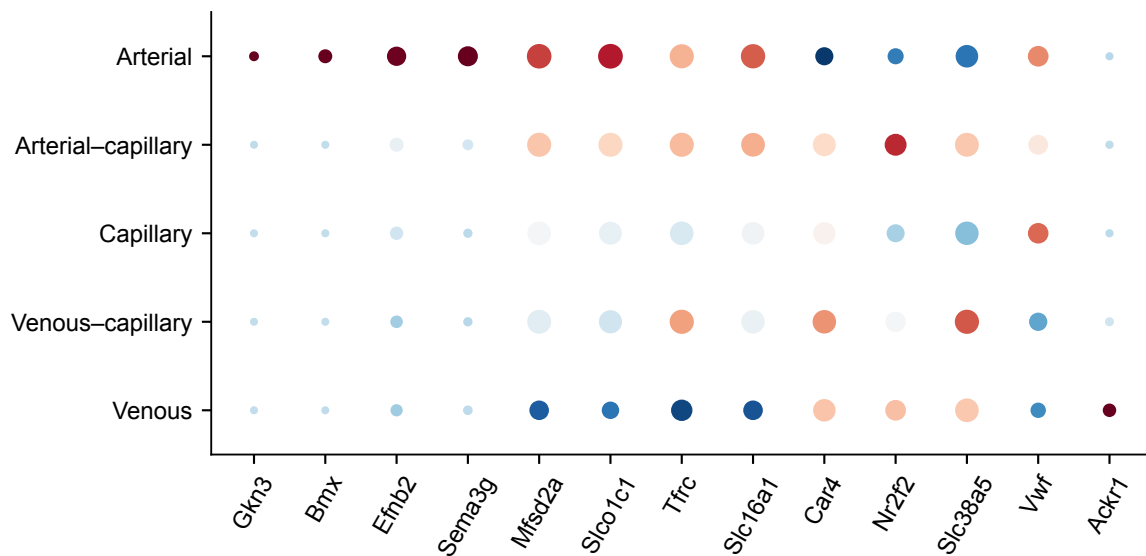
The ten numbered workbooks and full-results CSV parts of Table S9 are provided directly in ESM_2.zip (Supplementary Tables) without subfolders. CSV files use UTF-8 encoding. Blank or NA entries mean unavailable, not zero. Original reported and reanalyzed results are distinguished. Cell-level, pooled-library and clonal-culture statistics do not establish independent patient-level evidence.

a Cell retention by mouse**b** Retained singlet QC**c** Retained singlet QC**d** Singlet atlas**e** Analysis subsets

Colour: gene-wise z score (-2 to +2); size: detection fraction

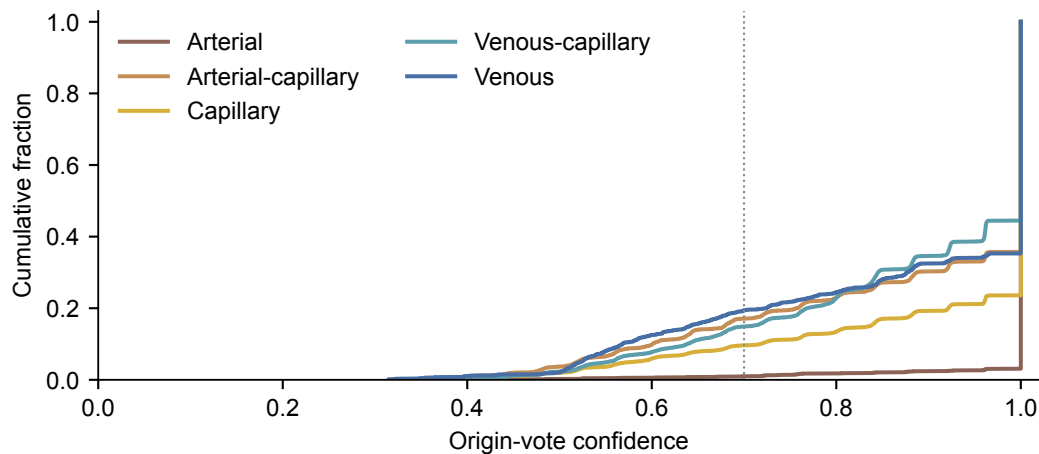
a

Vascular-origin marker profiles



b

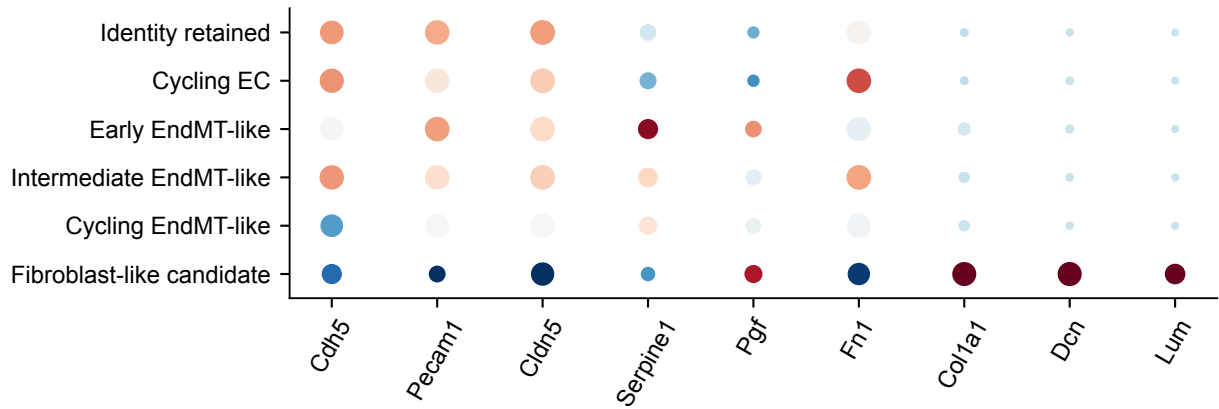
Confidence across inferred origins



Colour: gene-wise z score (-2 to +2); size: detection fraction

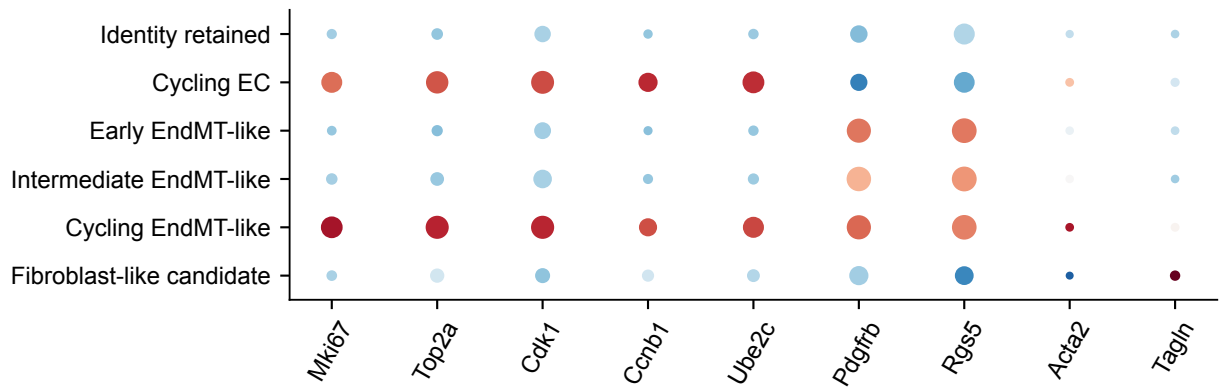
a

Endothelial and matrix markers



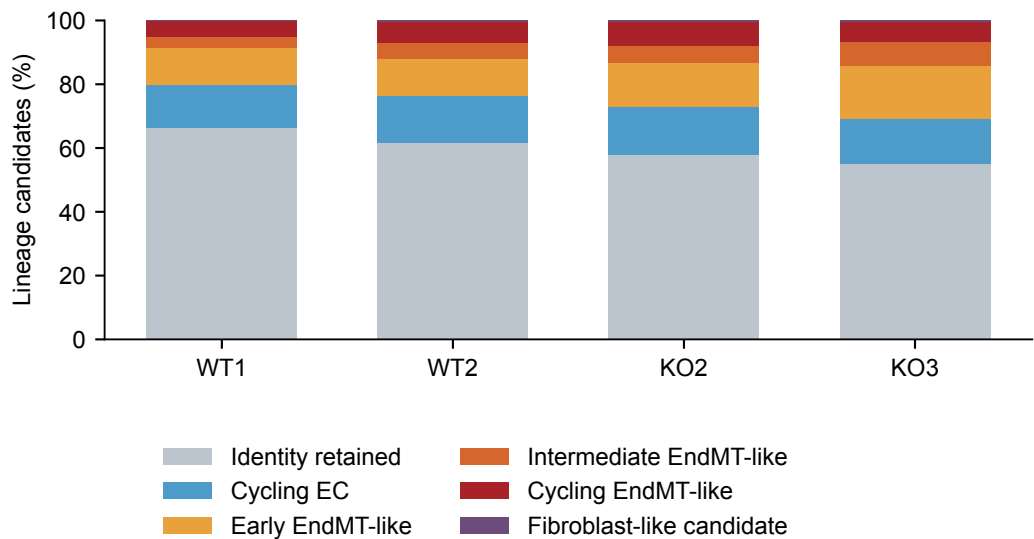
b

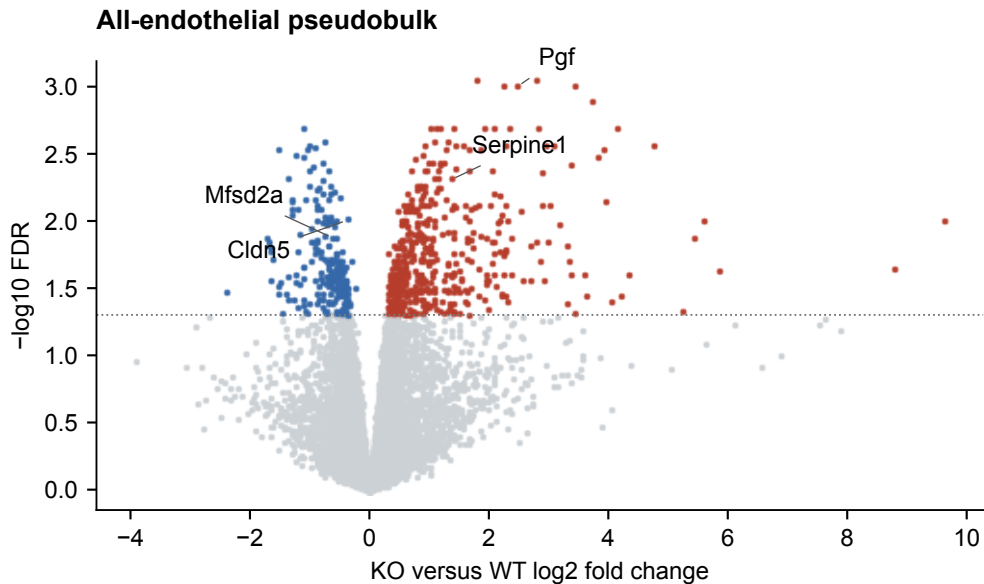
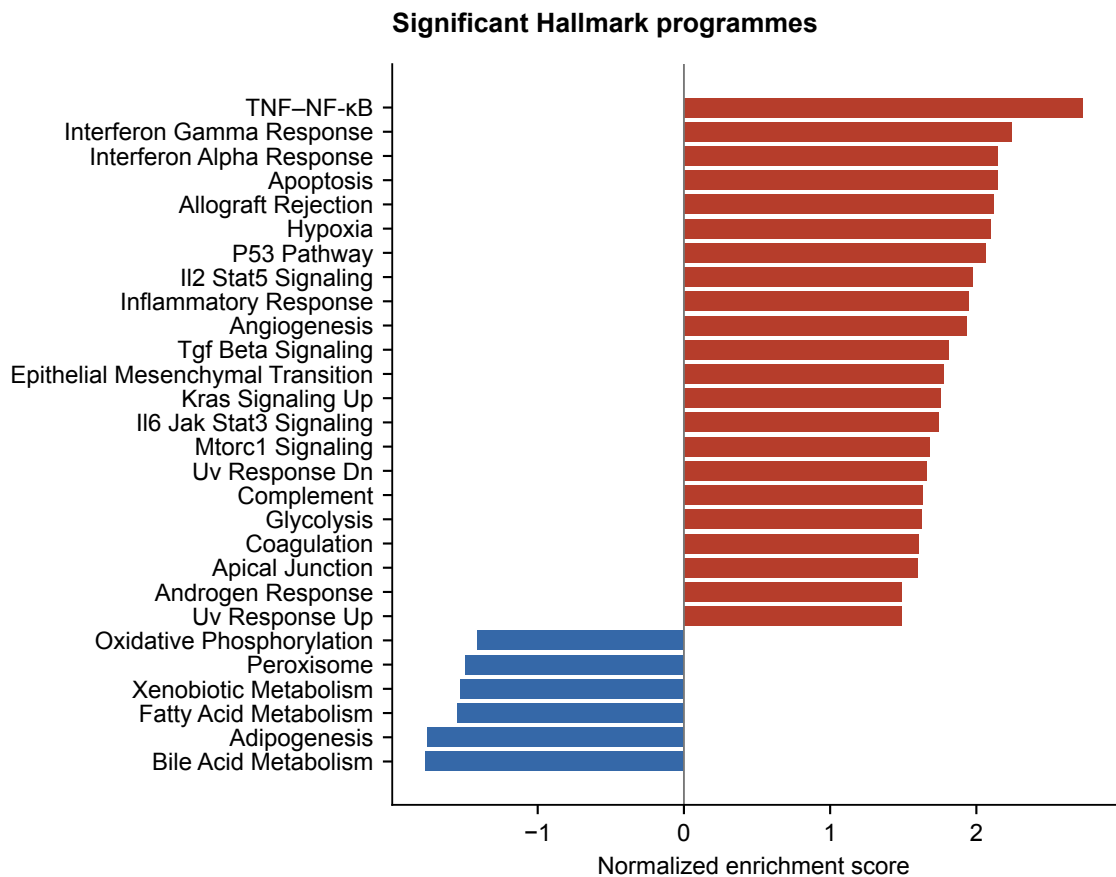
Proliferation and mural markers

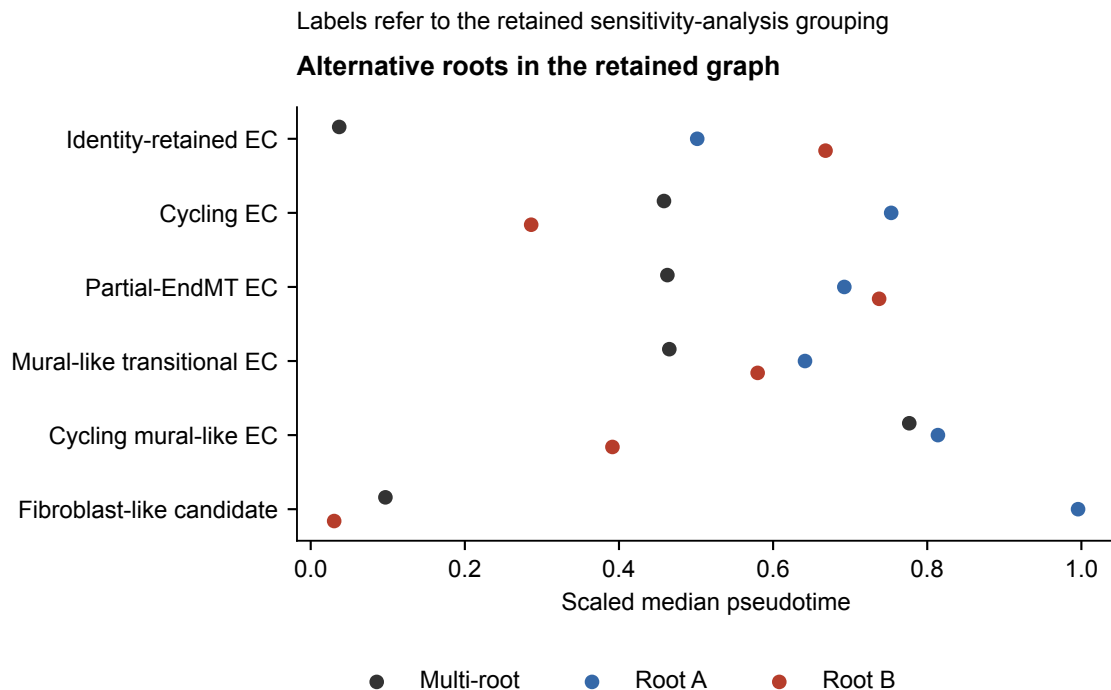
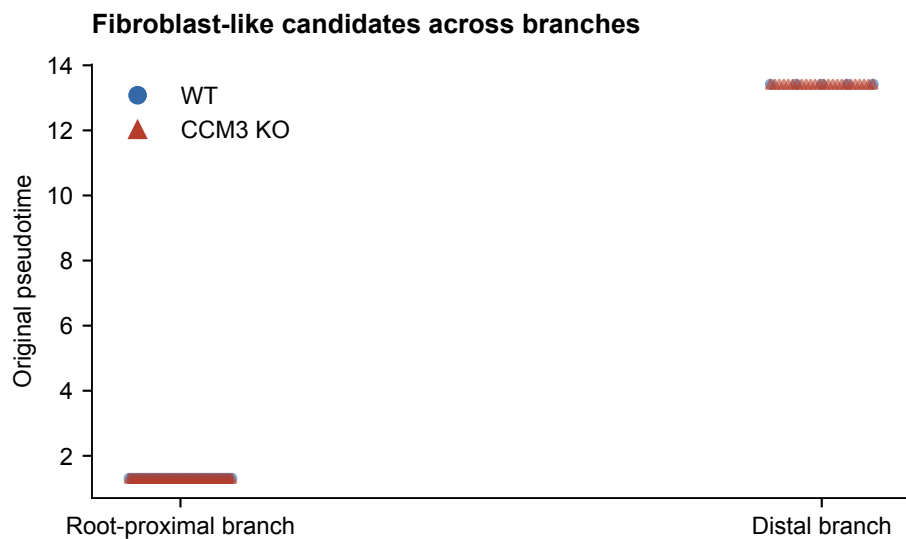


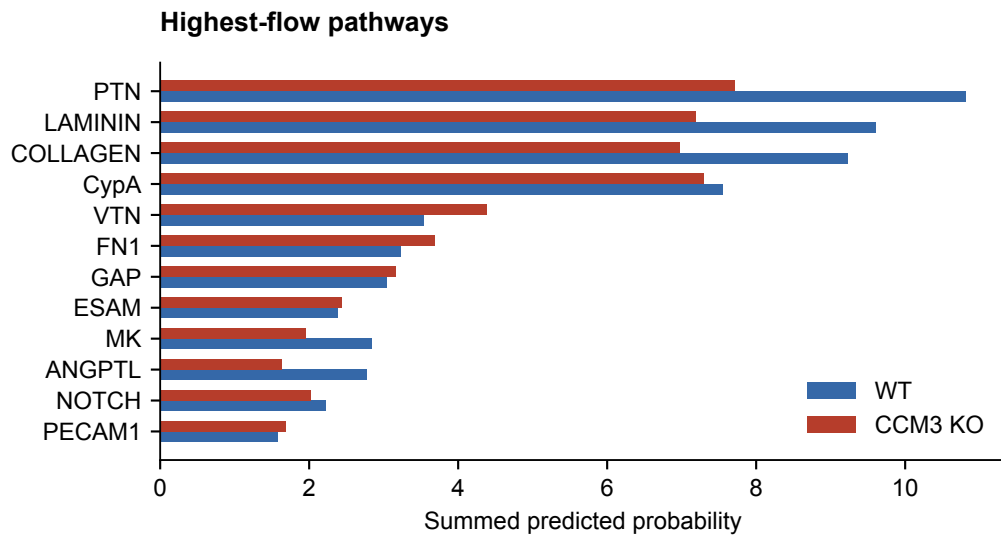
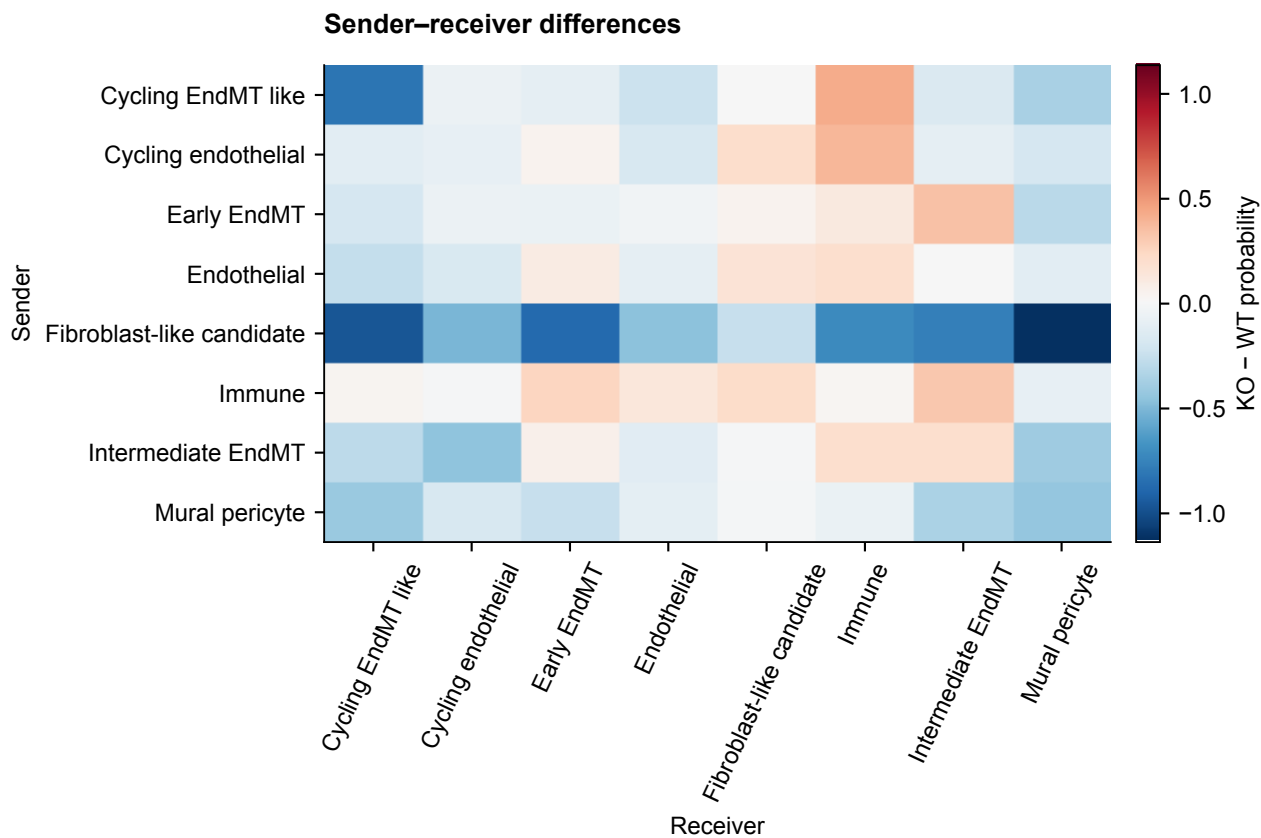
c

State composition in individual mice



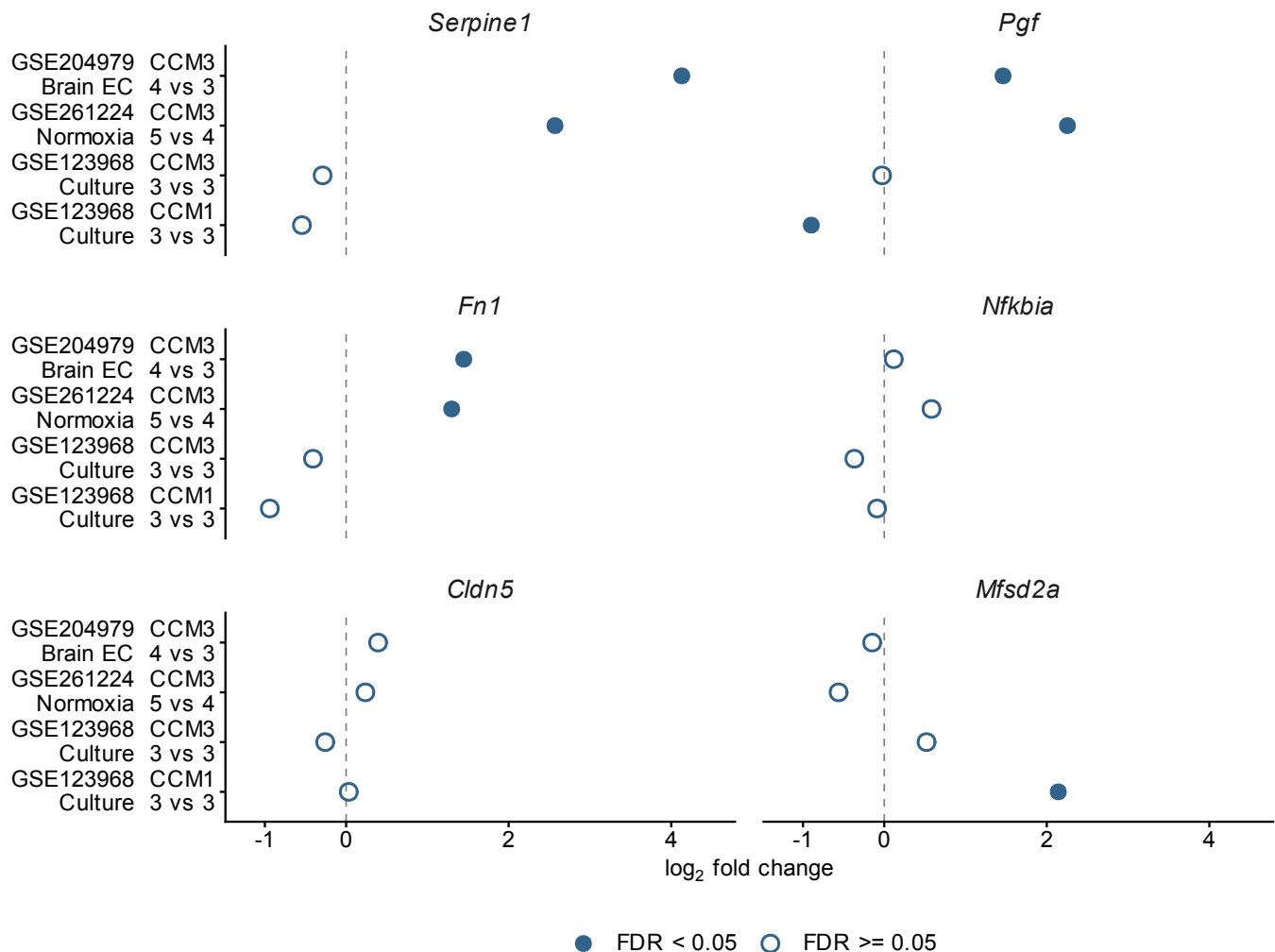
a**b**

a**b**

a**b**

Candidate gene effects in additional endothelial datasets

Loss versus matched control; sample counts are loss vs control



Separate models and harvest times; differences do not establish genotype severity.