An integrated single cell and spatial omics atlas of human prenatal development

Authors: Simone Webb†[1,2], Antony Rose†[1,2], Chuan Xu†[3], Lloyd Steele†[2], Masami A Kuri†[2], Emily Stephenson†[1,2], Kemal Inecik†[4], Daniyal Jafree†[2,5], April R Foster[2], Daniela Basurto-Lozada[1,2], Nana-Jane Chipampe[2], Anna V Pournara[2], Marc-Antoine Jacques[6,7], Ken To[2,3,8], Chloe Admane[1,2], Efpraxia Kritikaki[1,2], Marta A Chroscik[1,2], David Horsfall[1,2], Julia Foreman[6], Koen Rademaker[2], Julia Karjalainen[9], Anna Laddach[10], Shaista Madad[2], John E G Lawrence[2,8], Vitalii Kleshchevnikov[2], Steven Lisgo[1], Jimmy Tsz Hang Lee[2], Joséphine Blevinal[11], Ahlam Alqahtani[1], Stanislaw Makarchuk[2], Jake Jackson[1], Ekin Ucuncu[5], Teresa P Silva[5], Valentina Lorenzi[2,6], Fereshteh Torabi[2], Rachel A Botting[1], Kenny Roberts[2], Bayanne Olabi[1], Keerthi P Chakala[2], Leander Dony[4,12], Gabriele Dall'Aglio[4], Ana-Maria Cujba[3], Holly J Whitfield[2], Christine Hale[2], Justin Engelbert[1], Tong Li[2], Elena Prigmore[2], Nusayhah Hudaa Gopee[1,2], Michael W Mather[1], Ben Talks[1], Lijiang Fei[3], Diana Pereira[13], Corina Moldovan[14], Rasa Elmentaite[3], Krzysztof Polanski[3], Alexander V Predeus[2], Pasha Mazin[2], Vicky Rowe[2], Semih Bayraktar[3], Vincent R Knight-Schrijver[3], Menna Clatworthy[15], Kazumasa Kanemaru[3], James Cranley[3,16], Andrew Guo[15], Sanjay Sinha[3], David McDonald[1], Andrew Filby[1], Janet Kerwin[1], Peng He[17,18], Inês Sequeira[13], Roser Vento-Tormo[2], Malte D. Luecken[4,19], Alain Chédotal[^11,20,21], Deanne Taylor[^22,23], Silvia Martin-Almedina[24], Pia Ostergaard[24], Sam Behjati[2,25], Elizabeth Robertson[26], Nicola K Wilson[3,7], John Marioni[5,27], Helen V Firth[2,28], Caroline F Wright[29], Olivier Pourquie[30], Laure Gambardella[2], Paula Alexandre[5], Berthold Gottgens[3,7], Kerstin B Meyer[2], MS Vijayabaskar[2], Mo Lotfollahi[2], Sarah A Teichmann*[3,16,31], Fabian J Theis*[4], Fränze Progatzky*[9], Omer Ali Bayraktar*[2], Muzlifah Haniffa*[^2,28,32]

† co-first authors

* co-corresponding authors

^ on behalf of the Human Cell Atlas Development Biological Network

Affiliations:

[1]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK

[2]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

29 [3]Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus,
30 University of Cambridge, Cambridge, UK
31 [4]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Neuherberg,
32 Germany
33 [5]Developmental Biology & Cancer Department, University College London (UCL) Great Ormond Street
34 Institute of Child Health, UCL, London, UK
35 [6]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome
36 Genome Campus, Cambridge, UK
37 [7]Department of Haematology, University of Cambridge, Cambridge, UK
38 [8]Department of Surgery, University of Cambridge, Cambridge, UK
39 [9]Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK
40 [10]Nervous System Development and Homeostasis Laboratory, the Francis Crick Institute, London, UK
41 [11]Sorbonne Université, INSERM, CNRS, Institut de la Vision, Paris, France
42 [12]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
43 [13]Center for Oral Immunobiology and Regenerative Medicine, Barts Centre for Squamous Cancer, Institute
44 of Dentistry, Queen Mary University of London, London, UK
45 [14]Department of Pathology, Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
46 [15]Molecular Immunity Unit, Department of Medicine, University of Cambridge, Cambridge, UK
47 [16]Department of Medicine, University of Cambridge, Cambridge, UK
48 [17]Department of Pathology, University of California, San Francisco, CA, USA
49 [18]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California,
50 San Francisco, San Francisco, CA, USA
51 [19]Institute of Lung Health and Immunity (LHI), Helmholtz Munich, Comprehensive Pneumology Center
52 (CPC-M), Germany; Member of the German Center for Lung Research (DZL)
53 [20]Institut de Pathologie, Groupe Hospitalier Est, Hospices Civils de Lyon, Lyon, France
54 [21]University Claude Bernard Lyon 1, MeLiS, CNRS UMR5284, INSERM U1314, 69008, Lyon, France
55 [22]Department of Biomedical and Health Informatics (DBHi), The Children's Hospital of Philadelphia,
56 Philadelphia, PA, USA

57 [23]Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA,
58 USA
59 [24]School of Health and Medical Sciences, City St George's University of London, London, UK
60 [25]Department of Paediatrics, University of Cambridge, Cambridge, UK
61 [26]Sir William Dunn School of Pathology, University of Oxford, Oxford, UK
62 [27]Cancer Research UK, Cambridge Institute, University of Cambridge, Cambridge, UK
63 [28]Cambridge University Hospitals Foundation Trust, Addenbrooke's Hospital, Cambridge, UK
64 [29]Department of Clinical and Biomedical Sciences, St Luke's Campus, University of Exeter, Exeter, UK
65 [30]Department of Genetics, Harvard Medical School and Department of Pathology, Brigham and Women's
66 Hospital, Boston, MA, USA
67 [31]CIFAR Macmillan Multi-scale Human Programme, CIFAR, Toronto, Canada
68 [32]Cambridge Institute for Therapeutic Immunology and Infectious Diseases (CIITID), University of
69 Cambridge, UK

Abstract

Single cell genomics has enabled analysis of human prenatal development at unprecedented resolution. However, most studies have relied on dissociated tissues during restricted windows of development, limiting insights into how spatially distributed networks of cells, and multicellular niches emerge and adapt to distinct organ microenvironments *in situ.* Moreover, existing human developmental atlases have not yet been harmonised, and we thus lack a comprehensive catalogue of known cell types in the developing human body.

Here, we introduce the Human Developmental Cell Atlas (HDCA), a unified structural, cellular and molecular resource for prenatal human development. The HDCA integrates published and unpublished single cell/nucleus RNAseq atlases across prenatal organs, and includes a newly generated, spatially resolved, multimodal cell atlas of intact human embryos. Spanning 4-22 post conceptional weeks, capturing embryonic and early to mid fetal stages, the HDCA contains ~4.6 million cells/nuclei which resolve into ~450 cell types, explorable with a bespoke web portal.

For a global overview of the human embryo's multicellular communities, we applied unsupervised deep learning to our intact human embryo spatial data, charting 114 tissue niches that are structural and signalling hubs for the cellular interactions of the embryo. Guided by these niches, we profiled cellular networks over space and time, not examinable using single-organ atlases. In so doing, we revealed tissue-specific fibroblast patterning from previously undescribed mesenchyme progenitors, early diversification of organ-specific blood capillaries and lymphatic vasculature, emergence of neural crest cell fates, the formation of placode- and neural crest-derived peripheral sensory neurons, and how tissue niches guide peripheral neuron maturation and axonal migration. The HDCA thus serves as a comprehensive step towards a comprehensive understanding of human prenatal development, and a template towards unravelling the biology of congenital disorders.

Introduction

Human development is complex, spanning fertilisation, blastocyst formation, germ layer emergence, and the coordinated development and maturation of organs, all underpinned by expansive diversification of cell lineages from the initial zygote. *In vivo* studies in model organisms have uncovered key molecular mechanisms guiding developmental events, and together with advances in human stem cell biology, have furthered *in vitro* experimental culture systems that can replicate some aspects of early embryogenesis and human tissue development. However, *in vitro* culture systems and model organisms often do not fully faithfully recapitulate processes in humans. Studies of human embryos[1–4] have provided anatomical-level understanding of development and demonstrated that congenital disorders arise from anomalies in early developmental processes. However, anatomical information is poorly integrated with high resolution multimodal cellular and molecular information, which is required to thoroughly understand how human development is orchestrated across tissues in anatomical space and gestational time.

Single cell and spatial genomic technologies have transformed our ability to profile human development at unprecedented resolution. Initiatives, such as the Human Cell Atlas (HCA) Development Bionetwork[5], have mapped many organs, mainly across the first and second trimester of development, identifying temporal dynamics and cell types, e.g., in the brain[6–9], gut[7,10,11], heart[7,12–14], yolk sac[15], liver[7,16], bone marrow[17], kidney[7,16,18,19], lung[20], skeleton[21–23], skin[24,25], placenta[26] and thymus[27]. These embryonic and fetal reference atlases have been applied to clinical research, improving our understanding of altered development[17], the origin of childhood cancers[19,28], and inflammatory skin diseases[25]. However, the majority of this research profiled individual organs from different donors and with bespoke sampling protocols per lab, posing a challenge for understanding cell type presence and function across the human body. Leveraging large scale computational biology and the infrastructure and principles of the HCA Bionetworks, a new wave of integrative studies have overcome this barrier within tissues (e.g., the adult lung[29]), across complex tissue networks (e.g., developmental[24] and adult[30] immune systems), and across the whole body [31,32], powering advances in rare cell characterisation, addressing inconsistencies in cell type nomenclature across labs, and even creating frameworks to identify gene programme-disease links[33]. However, integrated and cross-tissue approaches for human development remain limited due to tissue restrictions and scarcity. Consequently, major distributed tissue networks in development such as the stroma, vasculature and peripheral nervous

system (PNS) remain under-explored, and given their complex architecture, would gain much from spatial profiling. Whole organism studies can address this challenge, as seen in model organisms[34–38], but human studies are limited and focus only on very early time points[39–42]. Biological insights have thus remained largely tissue-restricted, and the timing and anatomical location of gene expression implicated in congenital disorders has remained incomplete.

Here, we present the Human Developmental Cell Atlas (HDCA, n=186, k>4.6 million cells/nuclei, 4–22PCW), a unified resource integrating community-generated, organ-focused sc/snRNAseq developmental datasets, alongside complementary single cell data of 6PCW human embryos we generated to fill a critical gap in late embryogenesis. We further produced spatial transcriptomics (Visium, Xenium) data of 6PCW human embryos, which combined with the HDCA, enabled holistic assessment of cellular dynamics across time and space and revealed key insights into: tissue microenvironment influences on early mesenchymal and fibroblast patterning, the timing of organotypic blood capillary endothelial cell specification, the establishment lymphatic vascular molecular heterogeneity, neural crest cell fates, the formation and origins of peripheral sensory neurons, and how tissue niches guide peripheral neuron maturation and axonal migration. Further, the HDCA data and models provide a foundation for future research integrating new developmental datasets, we hope even bridging the gap to the earliest implantation stages. To facilitate continued biological insights and developmental disorder research, and empower users to derive biological insights through their data, we developed an accessible webportal, Cherita, which integrates rare disease patient data from the Developmental Disorders Gene2Phenotype (DDG2P, https://www.ebi.ac.uk/gene2phenotype) dataset[43].

144 **Results**

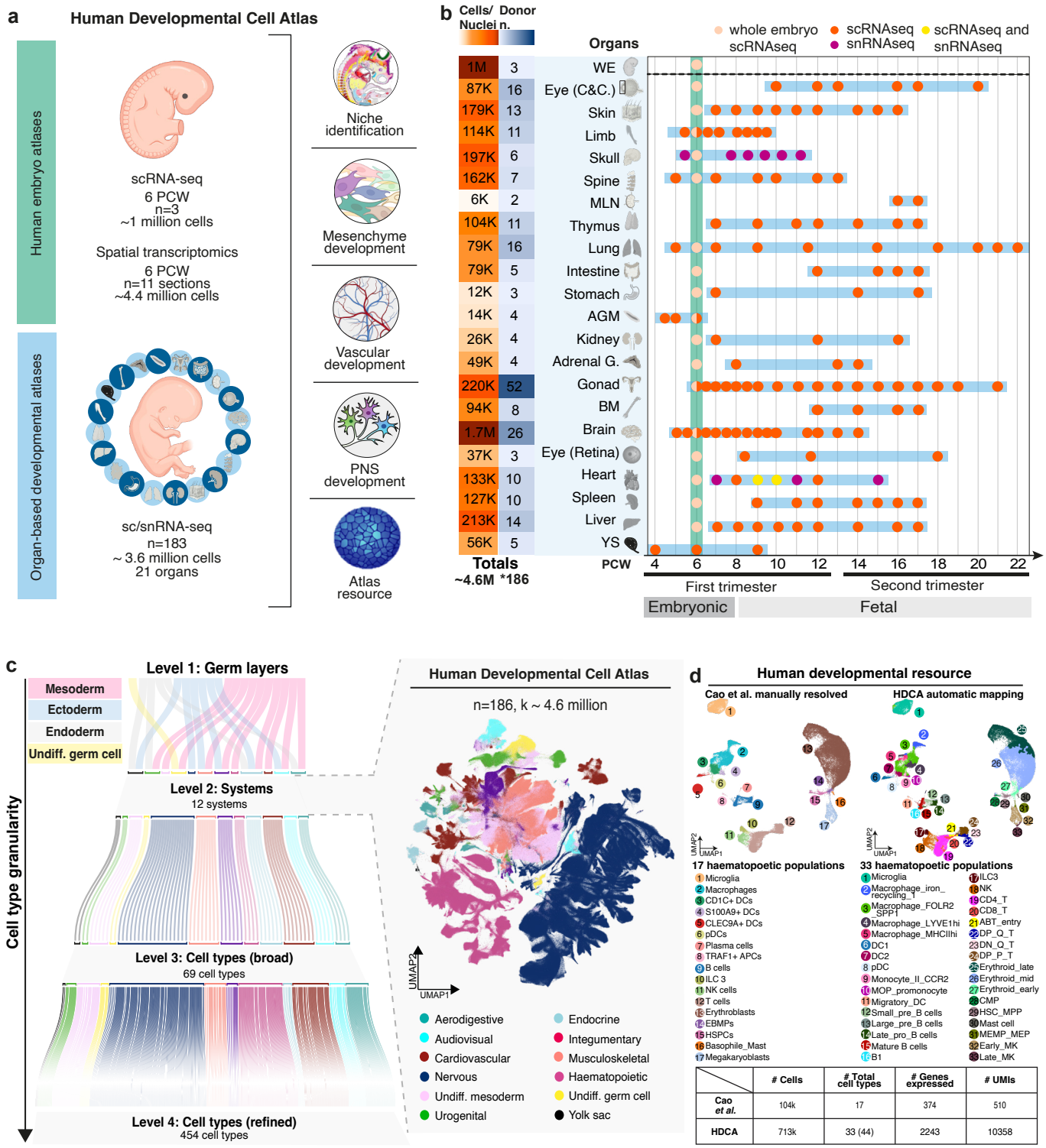
145 ***Integrated single cell and spatial omics atlas of human prenatal development***

146 To produce the HDCA, we first compiled published sc/snRNAseq datasets spanning 4-22 post-conception
147 weeks (PCW) and 21 developing organs (n=183 samples, k~3.6million cells) (**Fig. 1a-b; Supplementary**
148 **Table 1**). The selected datasets shared consistent sequencing technology (10x Genomics Chromium) and met
149 minimum QC standards for cellular abundance and cell type heterogeneity (see **Methods**). Some tissues and
150 organs were well sampled across gestation (e.g., brain), whereas others exhibited limited temporal span due
151 to tissue sampling (e.g., mesenteric lymph nodes only sampled 16-17PCW), or being restricted to a
152 developmental window (e.g., the yolk sac is absorbed at ~9PCW). As there was a relative scarcity of cells at
153 embryonic stages <8PCW, we generated complementary new data of 3 whole embryos (WE) dissected into
154 12 grid-like sections from late human embryogenesis at 6PCW (k~1million cells) (**Fig. 1a-c; Fig. S1a-h;**
155 **Supplementary Tables 2-4**). Our grid dissection approach for WE profiling enabled comprehensive sampling
156 across anatomical sites and spatially-informed cell annotation, leading to a powerful resource which allows
157 for comparison across the whole body within the same sample. scVI integration of the community
158 developmental datasets with the WE data produced the consolidated HDCA (n=186, k~4.6million), providing
159 a harmonized resource that complies with HCA metadata standards (https://data.humancellatlas.org/metadata)
160 (**Fig. S1i; Fig. S2a-c; Supplementary Table 5-8**).

161

162 After harmonising the cell type annotation labels provided in the public data, we performed expert manual
163 community-driven cell type annotations to: i) incorporate the newly integrated WE data, and ii) further resolve
164 cell types with homogenous labels in organ atlases due to their not being the study focus and assigned a generic
165 label (e.g., Mesenchyme), or undersampled (e.g., Glia) in public data (see **Methods**). This strategy resulted in
166 4 tiers of annotations in the final HDCA, capturing: germ layers (Level 1), 12 biological systems (Level 2),
167 69 broad cell types (Level 3), and 454 refined cell types (Level 4) (**Fig. 1c; Fig. S2d; Supplementary Table**
168 **9-10**). Here, refined cell types, e.g., 'CD4 T cell', were grouped into broad cell types which corresponded
169 generally to lineage, e.g., 'lymphoid'. Our systems-level annotations included: aerodigestive, audiovisual,
170 cardiovascular, endocrine, haematopoietic, integumentary, musculoskeletal, nervous, undifferentiated germ
171 cell, undifferentiated mesoderm, urogenital, and yolk sac (**Fig. 1c**). Cell type frequency analysis confirmed
172 expected spatial cross-organ cell type specificity, with for example endoderm most abundant in the lung, gut

Figure 1: Integrated single cell and spatial omics atlas of human prenatal development



	# Cells	# Total cell types	# Genes expressed	# UMIs
Cao *et al.*	104k	17	374	510
HDCA	713k	33 (44)	2243	10358

Figure 1: Integrated single cell and spatial omics atlas of human prenatal development

a) Schematic of data acquisition and overview of the Human Developmental Cell Atlas (HDCA). Top: *De novo* whole human embryo scRNA-seq data (n=3 samples; k>1million cells) and spatial transcriptomics (left: Visium, right: Xenium 5000-plex, 11 sections, k~4.4million cells). Bottom: Community acquired sc/snRNA-seq datasets of developmental tissues (n=188; k>3.6 million; 21 organs). **b)** Graphical illustration showcasing the summary of all single cell and nuclei data included in the HDCA atlas. Dot colour indicates data chemistry and public vs *de novo* whole human embryo data. For visualisation purposes, if an age is between 2 integers, a single dot is shown. All ages are outlined in **Supplementary Table 6**. Individual organs were grouped using a hierarchical dendrogram derived from scVI of public data only (as outlined in **Supplementary Table 10**). WE = Whole Embryo, C & C. = Cornea and Conjunctiva, MLN = Mesenteric lymph node, AGM = Aorta-gonad-mesonephros, Adrenal G. = Adrenal gland, BM = Bone marrow, YS = Yolk sac. Black dotted line separates whole human embryo data from public collected data. The first column on the left hand side indicates the number of data points for each organ while the 2nd column indicates the number of tissue samples per organ. The total number of data points (k>4.6million) and individual tissue samples (n=186 total, some contribute to more than 1 organ. **c)** Left: Sankey illustration depicting the level of annotations within the HDCA data increasing in granularity going top down. Right: UMAP visualisation of HDCA data (LVL2 system, n=186; k>4.6million). **d)** UMAPs of hematopoietic cells from Cao et al[47], colored by author-refined annotations (left) and by cell type labels transferred using the HDCA reference mapping model (right). Tables below summarize the number of cells, total annotated cell types (the number in parentheses denotes the total number of cell types present in the HDCA), and average expressed genes and UMI counts between the two datasets.

173 and stomach; ectoderm in brain and retina; and germ cells in gonads (**Fig. S2d**). Visualising the HDCA across
174 time showed that ectodermal neurons and glia were well-captured in the early WE 6PCW data, and
175 mesodermal stromal cell types were present across most public organ datasets across time (**Fig. 1c; Fig. S1i**).
176

177 ***Benchmarking atlas integration with trajectory fidelity***

178 The HDCA spans 6-22 PCW and thus gives a unique opportunity to model how cellular identities transition
179 over time. However, modelling time in single cell data is complex, and conventional benchmarking metrics
180 such as scIB[44] for biological conservation and batch correction, do not specifically factor for this (**Fig. S3a**).
181 We therefore developed a custom benchmarking framework focused on assessing how well any identifiable
182 cellular trajectories in the embedding conform to community-consensus ground truth: single cell trajectory
183 representation metrics (scTRAM)[45], which builds on existing scIB[44] metrics by additionally assessing if
184 differentiation trajectories remain conserved following integration. scTRAM evaluates trajectory fidelity
185 along 3 axes: topology conservation (whether connectivity and branching match the reference graph), cell-
186 order continuity (whether cells are arranged in the correct progression along directed edges) and trajectory
187 dominance (whether the lineage manifold is preserved and not overwhelmed by batch or other nuisance
188 structure) (**Fig. S3a**). The HDCA public data were pre-processed and integrated using commonly used tools
189 for scRNAseq integration including scVI, scPoli, TarDis, Harmony, Scanorama and PCA, after which
190 embeddings were assessed using both scIB and scTRAM. scVI attained the highest combined score (0.494),
191 followed by scPoli (0.437) and TarDis (0.421), with PCA, Harmony and Scanorama performing sub-optimally
192 (**Fig. S3b**). scVI also achieved the highest mean scTRAM score when stratified by trajectory group (**Fig. S3c**).
193

194 To assess performance differences along biologically meaningful developmental paths, we computed per-
195 edge scTRAM scores across the three curated ground-truth lineages (germline, haematopoietic, and peripheral
196 nervous system), and tested for significant differences between scVI and scPoli (**Fig. S3d**). scVI showed
197 higher scores on the majority of informative edges, including late haematopoietic branches, underscoring that
198 trajectory preservation is graded rather than binary and that integration methods introduce systematic biases
199 at specific branch points. For the haematopoietic lineage, UMAP visualisation with scVI exhibited a
200 continuous, correctly oriented path from HSC/MPP and LMPP/MLP through pro-B and pre-B
201 intermediates to mature, cycling and plasma B cells, whereas the scPoli embedding showed less

202 consistent ordering and locally misdirected flow, particularly around late pro-B and immature
203 compartments (**Fig. S3e**). UMAP visualisation of the germline lineage also implies that the atlas yielded
204 a more continuous progression from primordial germ cells through oogonia and pre-oocyte to oocyte and
205 better resolved branching towards pre-spermatogonia and granulosa cells, with smoother connectivity
206 and correctly oriented CellRank-inferred directional flow relative to the García component (**Fig. S3f**).
207 Together, these findings support the selection of scVI for downstream analyses as it provided the strongest
208 balance between trajectory fidelity and batch integration across the developmental atlas.
209
210 ***A comprehensive, adaptable human developmental resource***
211 To ensure the continued utility of the HDCA as a reusable reference, we provide models to the wider
212 community for data integration (scVI), cell type label transfer (CellTypist, scArches), and niche transfer
213 (NicheCompass) (**Fig. S4a**). We offer two complementary models for cell type label transfer: a deep learning-
214 based scArches[46] model that enables decentralised reference mapping through updateable, shared latent space,
215 and a CellTypist[30] model that performs logistic regression-based label transfer directly in gene expression
216 space. These models leverage the fine-grained HDCA annotations to support robust cell type label transfer
217 across different yet complementary analytical frameworks. Using the scArches model, we mapped two
218 independent human developmental datasets spanning different developmental windows[47,48] onto the HDCA,
219 demonstrating its value as a reference for interpreting users' data (**Fig. S4a**).
220
221 First, we mapped Cao et al[47], a large human mid-gestation cell atlas (10-18PCW; ~4.1 million cells and nuclei
222 generated using sci-RNA-seq3) onto the HDCA reference. Cao et al[47] described 657 cell clusters, yet the use
223 of clustering-like nomenclature hinders systematic alignment with standardized cell ontologies. We therefore
224 utilised their 77 higher-level groupings, which mapped to functional cell type annotations, for downstream
225 analyses. In spite of the substantially greater sequencing depth of the HDCA, we observed robust alignment
226 of cells across datasets within the shared latent space (**Fig. S4b**), and cell type label transfer using both
227 scArches and CellTypist models converged on highly consistent predictions (**Fig. S5a-b**).
228
229 Concordant cell types between Cao et al[47] and the HDCA included erythroid, myeloid, and lymphoid
230 populations, as well as multiple cell types in the vascular, muscle, epithelial, and neural lineages (**Fig. S4b-c;**

231 **Fig. S5a-b**). Within these shared groups (72 out of the 77 broad cell type groups, excluding placental lineages),
232 the HDCA reference was able to provide increased annotation granularity for 44 cell types (**Supplementary**
233 **Table 11**). Cao et al[47] further reclustered and reannotated three major cell compartments, including
234 hematopoietic (~104k cells), endothelial (~89k), and epithelial (~282k), to achieve higher-resolution
235 classifications. We therefore assessed whether HDCA-based automated mapping could recapitulate and
236 potentially surpass these curated annotations. In haematopoietic populations, HDCA mapping recovered the
237 manually defined lymphoid, myeloid, erythroid, and progenitor compartments while further resolving
238 thymocyte states, macrophage subsets, and T and B cell subpopulations **(Fig. 1d**). Similarly, endothelial
239 populations were partitioned beyond broad vascular and lymphatic categories into molecularly and
240 functionally different subtypes, such as refined lymphatic endothelial subpopulations not delineated
241 previously (**Fig. S4d**). Epithelial lineages were likewise resolved into diverse tissue-specific populations (**Fig.**
242 **S4d**).
243
244 Where there was weaker cell type concordance, this generally reflected partial matching of related cell types,
245 differences in developmental timing, or disparities in sampling of rare cell states and tissues between the query
246 data and the HDCA reference (**Fig. S4b-c, S5a-b**). Reassuringly, cell types from tissues unique to the HDCA
247 (e.g., yolk sac), or uniquely sampled in Cao et al[47] (placenta) lacked reciprocal counterparts (**Fig. S4c;**
248 **Supplementary Table 11**). While the HDCA reference supports mapping of datasets from comparable
249 developmental stages, alignment to pre-implantation stages dataset[48] were limited to pluripotent-to-
250 endodermal axis and primordial germ cells and gonads, indicative of embryonic pluripotency programmes
251 persistent in germline populations (**Fig. S6**). Together, these underscore the need to generate comprehensive
252 *in vivo* references for early timepoints or to bridge this gap using *in vitro* embryoid models in the future.
253
254 ***An unsupervised systematic decomposition of the human embryo into niches***
255 The ability to identify cell types using single cell omics has provided enormous insights into human
256 development, and the application of spatial technologies enables their *in situ* location. However, cells do not
257 exist in isolation but form functional niches with their cellular neighbours, reflecting principles of self
258 organisation[49,50]. To gain insight into cellular niches in tissue, we generated spatial transcriptomics on sagittal
259 sections of an additional 6PCW whole embryo sample, including whole transcriptome (Visium) and 5,000

260 gene panel (Xenium) and multiplex protein imaging data (RareCyte) to map cell types and gene expression
261 patterns across space and tissue architectures (**Fig. S1f, S7a-e; Supplementary Table 1, 3, 5**). Through
262 collaboration with the Human Developmental Biology Resource (HDBR), macroscopic annotations of nearby
263 haematoxylin and eosin (H&E)-stained tissue sections were incorporated into the Xenium spatial
264 transcriptomic data, enabling multi-modal and multi-scale cross-referencing of histological features with
265 predicted cell type assignments and gene expression profiles (**Fig. S7a**). We then performed niche detection
266 on the Xenium data in an unsupervised manner using NicheCompass[51], a deep learning tool which generates
267 niche embeddings based on a cell's gene expression and that of its local neighbourhood at single cell resolution
268 (**Methods**).
269
270 ***Histologically discernible and novel tissue niches in the human embryo***
271 In total, we identified 114 niche clusters across 8 human embryo Xenium sections (3.1 million cells). We
272 annotated these niche clusters based on HDBR histopathological annotations of nearby section H&Es and
273 niche-level differentially expressed genes (DEGs) (**Fig. 2a-b; Fig. S8a-b; Supplementary Table 12**). 37
274 niches corresponded to focal anatomical structures, including large organs such as the heart, spine and liver,
275 as well as smaller structures such as the mesonephros, gonadal ridge, otic vesicle, notochord, pancreas, adrenal
276 gland and gallbladder **(Fig. 2b; Fig. S9a-b**). Among these focal structures, 32 niches corresponded to H&E
277 annotated features (which we will refer to as 'known'), while 5 focal niches were not distinguishable by H&E
278 annotation (hereafter, 'novel') (**Fig. 2a-c; Fig. S8a-b, S9a-b)**. 30 niches corresponded to the central nervous
279 system (CNS) (**Appendix A**). The other 47 niches featured across multiple anatomical structures, such as a
280 *Nerve* niche and several *Vascular-rich* niches, in keeping with their presence in distributed networks (**Fig. 2a-**
281 **c; Fig. S8a-b, S9a-b**). Niches also had varied presence across sagittal tissue sections, e.g., *Rathke's pouch*
282 and the *Gallbladder* were only captured in specific sections, highlighting the necessity of profiling multiple
283 sagittal sections in order to comprehensively capture distinct midline and lateral embryo structures (**Fig. S8c;**
284 **Supplementary Table 13**).
285
286 Within the known niches, we used DEG analysis to identify candidate drivers of organ formation, confirming
287 previous work with animal studies as well as revealing novel human markers (**Supplementary Table 12)**. We
288 identified various known drivers described in humans, such as *NR5A1* in *Adrenal glands* and *PAX2/8* in the

Figure 2: An unsupervised systematic decomposition of human embryos into niches

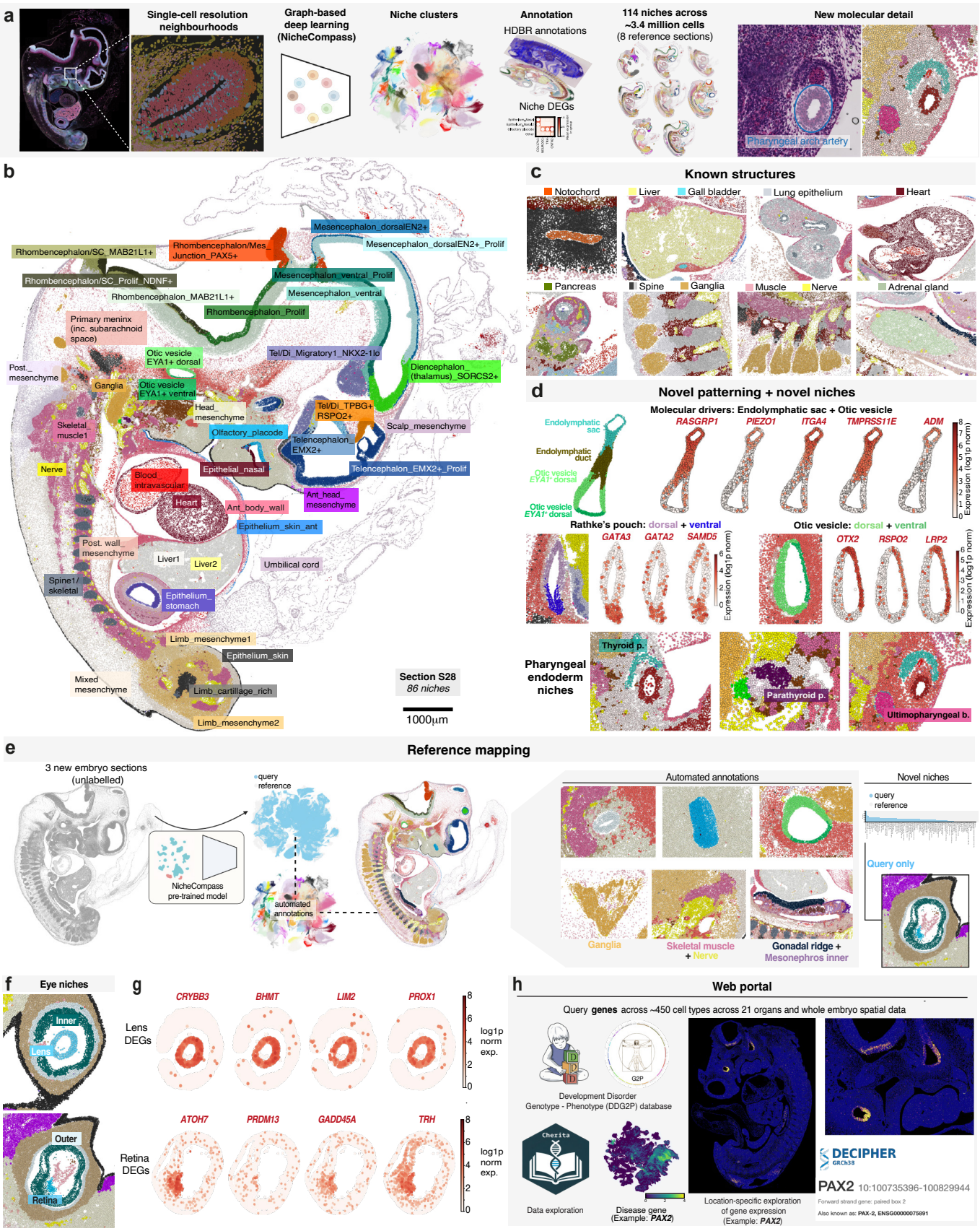


Figure 2: An unsupervised systematic decomposition of human embryos into niches

a) Overview of niche identification and annotation methodology. **b)** Xenium-5k profiling of human whole embryo section 28 (6PCW) with each cell coloured by niche assignment, with selected niche labels shown (see **Supplementary Table 13** for niche distribution across all sections).. Full Xenium NicheCompass niche labels shown in **Fig. S7c**. **c)** Focal regions of the whole embryo profiled by Xenium-5k coloured by niche across all embryo sections. We identified focal niches that corresponded to annotated structures from histology, such as the notochord and gallbladder. **d)** Top row: We captured the endolymphatic sac forming from the otic vesicle (S46 shown), including an *Endolymphatic_duct* niche. We queried differentially expressed genes between these structures. Log1p norm expression of example differentially expressed genes between the endolymphatic sac and otic vesicle. Middle row: Within focal niches that captured histological annotations, we identified dorsal-ventral patterning in Rathke's pouch (S69 shown) and the otic vesicle, (S133 shown) allowing us to query genes driving this difference. Log1p norm expression is shown for selected genes. D: dorsal, V: Ventral. Bottom row: We identified focal niches of pharyngeal endoderm that did not correspond to annotated histological structures beyond 'body mesenchyme' (sections S47, S93, and S103). D: dorsal, V: Ventral. **e)** Left: schematic of methodology for obtaining niche embeddings for 3 query sections. We trained a NicheCompass model using our reference data, and then fine-tuned this model with 3 new whole embryo sections. We clustered the k-nearest neighbour graph at high-resolution and named clusters based on the highest contribution from the reference data. To identify novel niches, we queried which clusters were only present in the query (see **Methods**). Section S133 shown. Right: Results from automated annotation, identifying focal niches e.g., the olfactory placode and otic vesicle (and a novel eye retinal niche) in the query without requiring expert histological annotations. **f)** Human eye structure coloured by niche, identifying outer, inner, lens, and retinal components, from sections S36 and S203. **g)** Example DEGs for the eye lens and retinal niches. **h**) Our bespoke Cherita web portal allows for interactive exploration of the HDCA across cell types, age, and genes. Cherita integrates developmental disorder gene information from the DDG2P[43] database (https://www.ebi.ac.uk/gene2phenotype)[192] and DECIPHER[193], a global platform for sharing rare disease patient data and interpreting variants, making the HDCA accessible to clinical and research communities. Information for *PAX2* can be viewed at https://www.deciphergenomics.org/gene/PAX2/expression; DECIPHER[193] currently lists seven sequence variants in this gene.

289 mesonephros and metanephros niches[52,53], and additionally investigated structures that have been poorly
290 studied. For instance, the endolymphatic sac, which maintains vestibular function by absorbing endolymph,
291 has never been captured in prior whole-embryo human transcriptomic analyses, and drivers of its development
292 are poorly understood. We identified the *Endolymphatic sac* as a distinct niche adjacent to the otic vesicle,
293 from which it is formed, as well as a novel *Endolymphatic duct* niche (**Fig. 2d, Fig. S10a-b**). We found that
294 *RASGRP1, ADM, ITGA4,* and *PIEZO1* were amongst the genes enriched within the endolymphatic sac (**Fig.
295 2d**; **Fig. S10b-d**). Profiling multiple sections allowed us to confirm these genes in a different tissue section
296 (**Fig. S10c**).
297
298 As well as identifying key potential drivers that control the formation of tissue and organ structures, our
299 approach also allowed us to uncover patterning within structures, including patterning of the *Rathke's pouch*
300 and *Otic vesicle* (**Fig. 2d; Fig. S10a**). This captured known genes, such as *OTX2* expression as a known driver
301 for otic placode patterning[54], as well as unreported genes including *RSPO2* and *LRP2* (**Fig. 2d**). We revealed
302 finer resolution niche annotation for the mesonephros, demonstrating distinct glomerular
303 (*Mesonephros_inner_glomerular*; *REN*, *FGF1*, *MAFB1*, *CLDN5*) and tubular (*Mesonephros_inner_tubule*;
304 *PDZK1*, *BHMT*, *HNF4A*, *KCNJ15/16*) niches (**Fig. S11a-b**), and captured the *Olfactory placode* (*NEUROD1*,
305 *RPRM*, *TRH*, *CNTN2*) within nasal epithelium in select sections (**Fig. S11c-d**). Collectively, these analyses
306 illustrate how molecularly defined niches link established drivers of organogenesis to classical histological
307 structures, while also uncovering previously unrecognised spatial patterning during human embryonic
308 development.
309
310 We were also able to delineate 3 novel *EPCAM*$^{+}$*CDH1*$^{+}$ niches (implicated in mouse pharyngeal
311 endoderm[34,55,56]) located around the pharyngeal artery and with distinct gene expression profiles (**Fig. 2d**), but
312 which were only broadly annotated as body mesenchyme by H&E (**Fig. S12a**). The first niche was consistent
313 with a thyroid progenitor, being characterised by known genes for thyroid development such as *PAX8,*
314 *FOXE1,* and *NKX2-1* (**Fig. S12b**)[57]. The second niche was consistent with a parathyroid progenitor, being
315 characterised by *PTH* (parathyroid hormone) expression and genes required for parathyroid and/or thymus
316 development, such as *GATA3*, *MAFB, SIX1* and *EYA1*[5859]. Of note, the final pharyngeal endoderm niche
317 represented the transient ultimopharyngeal body (or ultimobranchial body) that gives rise to the calcitonin-

318 producing neuroendocrine cells (C-cells) of the thyroid[60,61]. This niche arises from the 4th pharyngeal pouch
319 in mammals and was present in a similar location to the parathyroid progenitor niche (**Fig. 2c**). C-cells have
320 been profiled in human development but only from 9PCW[62]. Our high-plex profiling of this niche revealed a
321 distinct transcriptomic profile for early embryonic C-cells compared to later timepoints (**Fig. S12b-d)**. This
322 niche expressed genes associated with immature embryonic C-cells in mice (*ISL1*[63], *FOXA1* and *FOXA2*[61],
323 *ASCL1*[60], *NKX2-1*[64]) and other genes associated with neuroendocrine differentiation (*HHEX*, *ENO2, DLL3*)[65]**.**
324 As expected, expression of more mature C-cell markers was low or absent. For instance, *CALCA*, encoding
325 the alpha isoform for calcitonin, was relatively lowly expressed in this niche, whereas *CALCB*, encoding the
326 beta isoform, was highly expressed (**Fig. S12b-d)**. We thus provide the first detailed characterization of early
327 pharyngeal endoderm niches in human development with spatial resolution, including the previously
328 uncharacterised ultimopharyngeal body. Moreover, this demonstrates that our niche-based approach can
329 resolve distinct molecular structures within tissues that are histologically similar.
330
331 Our NicheCompass approach incorporates anatomical information within niche labels, allowing for future
332 community automatic niche transfer and cell type assignment to new embryo tissue sections without domain
333 or histological expertise, even with a small number of query tissue sections. To demonstrate this approach,
334 we mapped a further 3 embryo sections to our core 8-section dataset using reference-query mapping,
335 identifying 90 niches in these new sections (**Fig. 2e; Fig. S13a-d; Methods**). Notably, we identified a retinal
336 niche (*Eye3_innermost_retina*) not present in the reference dataset (**Fig. 2f; Fig. S13**), as well as an
337 *NTN1*$^+$*WNT3*$^+$*GABRB2*$^+$ niche present within a focal region of *MEIS1*$^+$*HMX2*$^+$ otic vesicle (**Fig. S10a, c,**
338 **S13c**). *HMX2* is a marker of dorsolateral otic epithelium[66], and the focal *NTN1*$^+$*WNT3*$^+$*GABRB2*$^+$ niche
339 expressed genes that have been associated with semicircular canal development from the dorsolateral otic
340 vesicle in animal models (*WNT3*, *NR4A3*, *NTN1*)[67,68] (**Fig. S13c**), illustrating the granularity of our niche
341 capture. The human retina has not been studied in early embryogenesis, with prior transcriptomic studies
342 performed at later timepoints (8-9PCW) or in organoid models[69,70]. We showed that the embryonic retina was
343 characterised by expression of *ATOH7*, required for retinal ganglion cell (RGC) development[71]*; PRDM13*,
344 which specifies amacrine cell fate in mouse[72]; *GADD45A*, which characterises early neurogenic cells in mouse
345 retina[73]; and *TRH*, which encodes thyrotropin releasing hormone (**Fig. 2f-g; Fig. S13d**). *TRH* signalling has

346 been reported to specify cone subtypes in human retinal organoids[74], and thus we provide further evidence in
347 support of a role for *TRH* in the retina.
348
349 We finally wanted to confirm whether our niches would be informative for understanding the role of genes
350 implicated in developmental disorders. To illustrate the utility of niches captured, and after identifying the
351 retinal niche above, we investigated reported genes for congenital cataracts[75] (**Methods**), which is
352 characterised by lens opacity. Consistent with this pathology, congenital cataract genes were enriched
353 specifically in the *Eye1_lens* niche (**Fig. 2f-g; Fig. S13e**), including genes encoding crystallins (*CRYAA,*
354 *CRYBB3, CRYGS, CRYGC*), transcription factors (*MAF, FOXE3*), cytoskeletal proteins (*BFSP1*), and RNA
355 processing proteins (*TDRD7*)[75]. *TDRD7* was initially validated in developing mouse lens[76], but we illustrate
356 how this gene can now be easily interrogated in the context of human development through query of
357 expression across multiple embryo niches, demonstrating the potential value of our resource for querying
358 disease-associated genes given our comprehensive sagittal profiling and niche capture. Overall, we illustrate
359 how our data can serve as a resource for human embryo biology across diverse niches and generate
360 histologically-informed niche annotations for novel data.
361
362 In addition to providing light-weight access to the HDCA through the models described above, our bespoke
363 web portal, Cherita (https://cellatlas.io/studies/hdca/; private until publication), allows interactive exploration
364 of both droplet-based and embryo spatial datasets without requiring computational resource or technical
365 expertise (**Fig. 2h**). Cherita incorporates data from the DDG2P dataset
366 (https://www.ebi.ac.uk/gene2phenotype)[43], which shares detailed, evidence-based gene-disease models
367 curated from the literature. Given that our HDCA data covers 21 organs and the whole embryo, queried genes
368 can be investigated across different organs and developmental stages to contextualize spatiotemporal patterns.
369 For instance, loss-of-function variants in *PAX2* are known to cause Renal-Coloboma Syndrome [OMIM
370 #120330][77], characterised by optic nerve abnormalities, hypodysplastic kidneys, and, in some cases, high-
371 frequency sensorineural hearing loss. Consistent with this phenotype spectrum, we observed that *PAX2* was
372 highly expressed in the eye, metanephros, and otic vesicle niches (**Fig. 2h; Fig. S13f**). We thus provide the
373 first platform for tissue niche-level queries of human developmental tissues to dissect expression profiles

374 across structures, and for comprehensive investigation into developmental disorders linked to pathogenic
375 genetic variants.
376
377 ***Mesenchymal progenitors in the human embryo***
378 Leveraging the HDCA cross-temporal data and spatial niche identification, we then interrogated the major
379 distributed tissue networks in development which have, to date, have not been investigated in a cross-tissue
380 manner across human gestation: the developing mesenchyme, vasculature, and peripheral nervous system.
381
382 Within the developing mesenchyme, current understanding suggests that common fibroblast progenitors arise
383 from the mesoderm and neural crest and contribute to anatomical region-specific fibroblast subsets[78]. Single
384 cell transcriptomics technologies have elucidated novel heterogeneity in adult fibroblasts across tissues,
385 which previously was challenging to resolve due to unreliable surface markers[79]. However, fibroblast
386 diversity within the early stroma of the embryo remains poorly understood as fibroblasts have typically been
387 studied in an organ-specific manner. For instance, *HIC1*$^{+}$ and *FRZB*$^{+}$ fibroblast progenitors have been
388 described in the skeletal system and skin, respectively[22,80]. Despite recent whole organism profiling efforts[39–
389 42,81], the relatively small number of cells and single cell resolution spatial data means it remains unknown
390 how these different progenitor cells compare to each other and whether cross-organ populations.
391
392 From a joint integration of ~3.6 million predominantly mesenchymal cells (**Methods**) from both scRNA-seq
393 and Xenium data, we defined 108 cell types within 12 broad groups (**Fig. 3a**; **Fig. S14-16; Appendix B**).
394 Ten clusters were present at early timepoints (**Fig. 3b)** and remained undefined after expert annotation
395 (**Methods**), raising the possibility that they represented undefined progenitor populations. These clusters
396 expressed *FRZB*$^{+}$ +- *HIC1*$^{+}$, and other dermal fibroblast progenitor markers (**Fig. 3c**), indicating broad
397 expression of these previously reported genes for organ-specific progenitors. We then defined the spatial
398 distributions and niche residences of these 10 *FRZB*$^{+}$+-*HIC1*$^{+}$ fibroblast progenitor subsets across tissue
399 sections (**Fig. 3d; Fig. S14d, S17a**). The *PDGFRA*$^{+}$ progenitor population was non-specific and broadly
400 distributed. The remaining populations demonstrated more focal distributions, including enrichment in the
401 skeletal system (*RUNX2*$^{+}$), limb (*IRX3*$^{+}$*ISL1*$^{+}$), gut and anterior body (*FOXF1*$^{+}$), and head/jaw and limb
402 (*SATB2*$^{+}$*PAX3*$^{+}$ and *SATB2*$^{+}$*DCX*$^{+}$). We also delineated 4 *ZIC*^{hi} progenitor populations enriched in the scalp

Figure 3: Mesenchymal progenitor heterogeneity in the human embryo

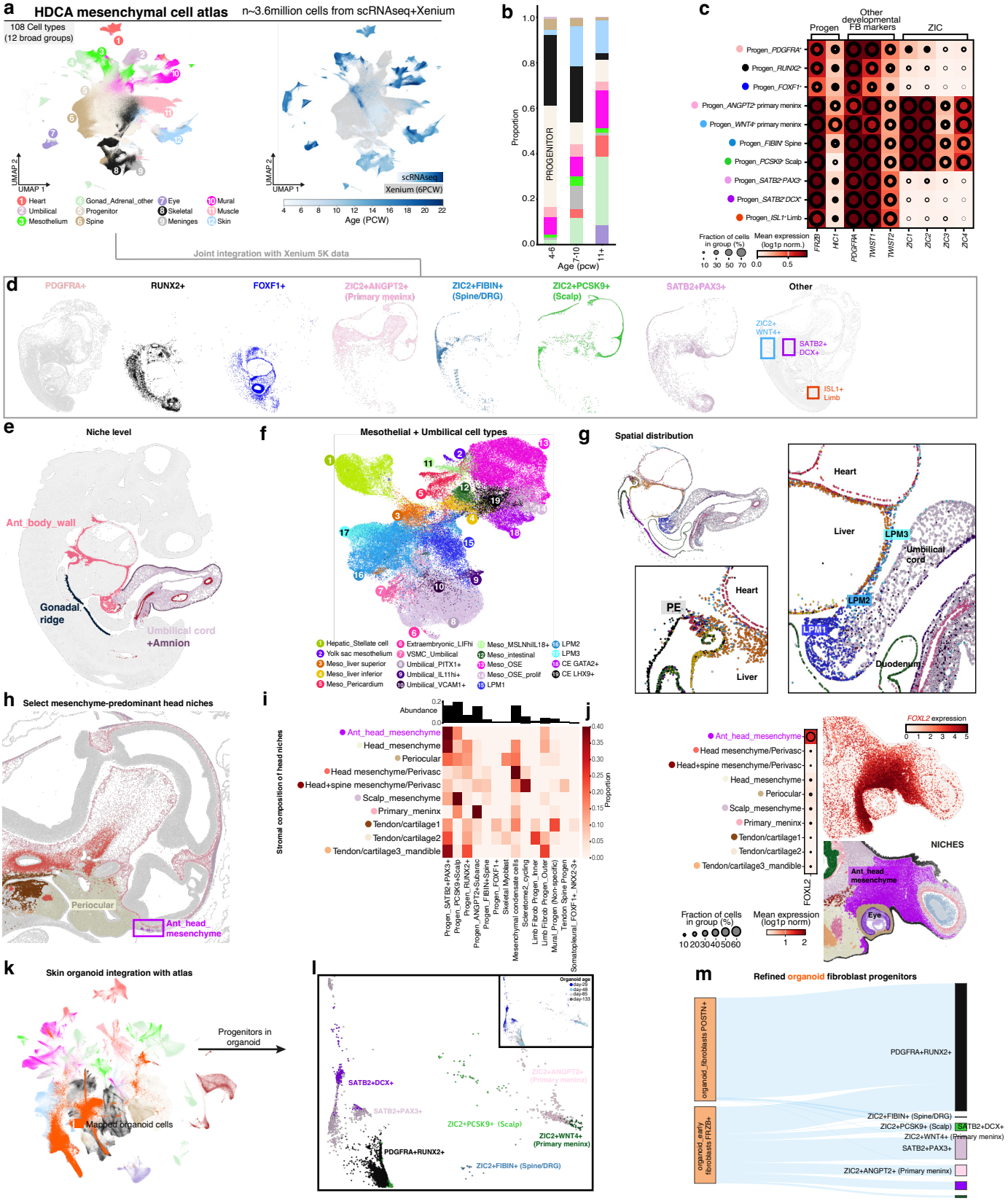


Figure 3: Mesenchymal progenitor heterogeneity in the human embryo

a) Left: UMAP embedding of predominantly mesenchymal clusters including scRNA-seq and Xenium-5k data, coloured by their broad cell type group. Across these broad groups we identified 108 cell types. Right: UMAP embedding coloured by age of sample. Xenium-5k samples (all 6 PCW) coloured in grey. PCW: post-conception weeks. **b)** Proportion of cells for each time grouping from each broad group, showing enrichment of progenitors in the early samples. scRNA-seq data only. **c)** Dotplot of expression of broad progenitor markers and *ZIC* genes in undefined human fibroblast progenitors in scRNA-seq data. **d)** Xenium-5k of whole embryo coloured by the undefined progenitor populations, showing distinct localisations (section S28, far right column of rarer populations in S47). **e)** Xenium-5k of whole embryo coloured by niches where mesothelial and umbilical cell populations are predominantly located (S46). **f)** UMAP embedding of mesothelial and umbilical cells coloured by cell type assignment. Hepatic stellate cells were located within the mesothelial group, consistent with a reported origin from septum transversum-derived mesothelium in mice. **g)** Location of mesothelial and umbilical populations in a Xenium-5k whole embryo section (S46). 2 broad clusters of mesothelial cells represented the LPM and PE. The LPM1 population was enriched in the anterior body mesenchyme lining, and LPM2 and LPM3 showed migration from this site. All LPM locations shown in **Fig. S18a**. PE: proepicardial organ. LPM: lateral plate mesoderm. **h)** Xenium-5k of whole embryo head coloured by niches with predominantly mesenchymal cells (S28). Other niches in this region (central nervous system) coloured in grey. **i)** Heatmap showing distribution of fibroblast subtypes by head mesenchymal niche across all Xenium 5k reference sections. **j)** Combined heatmap-dotplot of *FOXL2* expression by head mesenchymal niche. All niches shown in **Fig. S7c**. Example of lateral embryo section coloured by *FOXL2* expression (top), showing enrichment around the eye, and coloured by niche (S193; bottom). **k)** UMAP embedding of organoid cells (orange) compared to mesenchymal atlas coloured by broad cell type group. **l)** UMAP embedding of human organoid cells within the progenitor cluster, showing identification of Progen_SATB2+PAX3+, Progen_ZIC2+PCSK9+(Scalp), Progen_ZIC2+FIBIN+(Spine/DRG), and Progen_ZIC2+ANGPT2+ populations. Dotplots of these progenitor populations in the skin organoid using markers compared to human populations is shown in **Fig. S20b**. **m)** Sankey plot of re-assignment of original organoid labels, showing identification of fine fibroblast populations. Cell populations with a fraction of 0.02 or more are shown.

403 (*ZIC2*$^{+}$*PCSK9*$^{+}$), primary meninx (*ZIC2*$^{+}$*ANGPT2*$^{+}$, *ZIC2*$^{+}$*WNT4*$^{+}$), and the spine/posterior vertebral wall
404 (*ZIC2*$^{+}$*FIBIN*$^{+}$) (**Fig. 3c-d; Fig. S14d**). We corroborated these distinct locations through interrogation of the
405 anatomical origin of scRNA-seq data (**Fig. S17b**) and PAGA analysis (**Fig. S17c**). These progenitor clusters
406 also showed distinct gene expression signatures that were consistent in scRNA-seq and Xenium data (**Fig.**
407 **S14c**). Our analysis thus reveals novel regionally-patterned fibroblast progenitor populations that broadly
408 express previously used markers for organ-specific progenitors (*FRZB*, *HIC1*) but which can be
409 distinguished by other marker genes, redefining the stromal landscape in early embryos.
410
411 ***Lateral plate mesoderm heterogeneity and mesothelium formation***
412 Embryonic internal organs are embedded within coelomic cavities which are lined by epithelial monolayers
413 termed mesothelial cells, which arise from the lateral plate mesoderm (LPM)[82]. Notably, original author
414 annotations of the HDCA community data had not resolved LPM cells (**Fig. S15a-b**), possibly reflecting the
415 relatively small number of LPM captured when studied in an organ-specific manner, or that mesothelial
416 populations were discarded or lost during tissue dissection. Our spatial and integrative approach captured
417 coelomic cavities in the embryonic anterior body wall niche, including 3 LPM subtypes enriched in anterior
418 body mesenchyme (**Fig. 3e-f; Fig. S18a-c**). The *SFRP5*$^{+}$ LPM1 population was enriched in the anterior body
419 mesenchyme, particularly the inner lining of this structure (**Fig. 3g; Fig. S18a-b)**. Notably, *SFRP5* has been
420 reported to be a marker gene for an LPM progenitor in zebrafish[83]. The other 2 populations (LPM2, LPM3)
421 showed a more extensive distribution, with the *DKK1*$^{+}$*NRN1*$^{+}$ LPM3 population particularly observed in the
422 hepatic and cardiac regions (**Fig. 3g; Fig. S18a-b**), consistent with the LPM as a source of mesothelial
423 populations at these sites. In addition, we also observed a cluster of mesothelial cells located posterior to the
424 heart (**Fig. 3g**), consistent with the proepicardial organ.
425
426 Organ-specific mesothelium populations have been reported in prenatal lung[20], heart[14], liver[84], and
427 intestine[85], but similarly to fibroblast progenitors, it is unclear how these different populations are
428 transcriptionally similar or related, or the extent of their organ-specificity. Highlighting the power of our
429 atlas to resolve populations with fine (monolayer) distributions, we captured distinct mesothelial cell
430 populations lining the heart, intestine and inferior and superior liver. We also observed coelomic epithelium
431 subtypes lining the gonadal ridge and mesonephros **(Fig. 3e-g; Fig. S15d, S18c**). As we captured these

432 distinct mesothelial populations from across different organs, we were able to determine which previously
433 reported mesothelial markers for distinct organs were specific (e.g. *CALB2*, *SBSPON*)[86] and which were
434 shared (e.g., *TNNI*, *BNC1*, *PODXL, WT1*)[14,20] **Fig. S15c**). Additionally, we defined more comprehensive
435 gene signatures to distinguish these developmental mesothelial subsets (**Fig. S15c**). We thus
436 comprehensively decode the classical LPM structure with single cell spatial resolution, revealing novel LPM
437 subtypes with distinct spatial distributions, and profile the distinct mesothelial populations that they give rise
438 to across organs.
439
440 ***Extra-embryonic umbilical cord and mesenchymal cell heterogeneity***
441 Prior whole embryo analyses have not resolved extraembryonic mesenchymal populations[40,42], or have
442 grouped these populations as a single cluster[41]. Here, we revealed novel heterogeneity in the extraembryonic
443 tissue, capturing 4 different umbilical populations and an additional 2 extraembryonic mesenchymal
444 populations (**Fig. 3e**; **Fig. S16a-d)**. A *PITX1*$^{+}$ population was the predominant population in the umbilical
445 cord with a broad distribution, with the remaining populations showing more focal distributions. The
446 *SATB2*$^{+}$*IL11*^{hi} population was identified at the periphery of the umbilical cord. The *PITX1*$^{+}$*VCAM1*$^{+}$
447 population was enriched at the attachment to the embryo. We also profiled umbilical vascular smooth muscle
448 cells (*Umbilical_VSMC*) around the umbilical cord blood vessels (**Fig. S16c-d**), that have been ill-defined
449 to date[87]. The 2 extraembryonic mesenchymal populations, *Extraembryonic_CLIC3*^{hi} and
450 *Extraembryonic_LIF1*^{hi}, were at the periphery of the umbilical cord (**Fig. S16d**). These extraembryonic
451 populations expressed genes encoding *AFB* and *ALB*, associated with placental cells, but not trophoblast
452 markers (**Fig. S16e**). We thus capture multiple novel cell populations present within extraembryonic tissue
453 in the late embryo, including poorly defined supporting cells around umbilical vessels.
454
455
456 ***Decoding cranial embryo niches***
457 Elements of the vertebrate head arise from both mesodermal and neural crest sources[88]. Recent spot-level
458 spatial transcriptomic work has suggested distinct fibroblast subtypes in the head, but this study was limited
459 to only a focal region[40]. We identified 10 head niches formed predominantly of mesenchymal cells (**Fig. 3h;**
460 **Fig. S19a**). These mesenchymal head niches contained multiple fibroblast subtypes (**Fig. 3i**). In addition to

461 the non-specific *PDGFRA*$^{+}$ population, the most abundant progenitor subtypes in the head were the
462 *ZIC2*$^{+}$*PCSK9*$^{+}$ scalp, *ZIC2*$^{+}$*ANGPT2*$^{+}$ primary meninx, and *SATB2*$^{+}$*PAX3*$^{+}$ populations.
463
464 Notably, the *Ant_head_mesenchyme* niche showed a distinctive focal location in the anterior face, and was
465 markedly more abundant in association with the periocular niche (**Fig 3h; Fig. S19b**). The
466 *Ant_head_mesenchyme* niche predominantly contained *SATB2*$^{+}$*PAX3*$^{+}$ and *ZIC2*$^{+}$*PCSK9*$^{+}$ fibroblast
467 progenitors (**Fig 3i**), which were not unique to this niche. We therefore queried whether these fibroblasts
468 showed distinct gene expression in different niches. *FOXL2* expression was highly enriched in the
469 *Ant_head_mesenchyme* niche compared to other head niches (**Fig. 3j; Fig. S19c**).We therefore queried
470 *FOXL2* expression within *SATB2*$^{+}$*PAX3*$^{+}$ and *ZIC2*$^{+}$*PCSK9*$^{+}$ fibroblast progenitors across different niches.
471 We observed increased *FOXL2* expression with these progenitors in the *Ant_head_mesenchyme* niche
472 specifically (**Fig. S19d**), suggesting that these fibroblasts exhibit distinct gene expression across different
473 niches. Interestingly, *FOXL2* is implicated in eyelid formation, with loss-of-function mutations causing
474 eyelid abnormalities in blepharophimosis-ptosis-epicanthus inversus syndrome (BPES), in agreement with
475 the enrichment of the *Ant_head_mesenchyme* niche around the *Periocular* niche. Our analysis thus suggests
476 a distinct niche involved in the formation of the eyelid structure that is not identifiable from cell type
477 composition alone.
478
479 Overall, within the developing mesenchyme, we reveal novel complexity in fibroblast progenitors across the
480 embryo; comprehensively profile extraembryonic and mesothelial populations at single cell spatial
481 resolution, including migrating LPM; and identify and decode head mesoderm niches, revealing niches
482 contributing to the formation of specific structures, providing the most complete profiling of human
483 developmental fibroblasts to date.
484
485 ***Novel mesenchymal progenitors are recapitulated in human organoid model***
486 We aimed to determine if our newly-defined human fibroblast progenitor populations were captured and
487 corroborated by an induced pluripotent stem cell (iPS)-derived organoid model containing mesenchymal
488 lineage cells. To this end, we mapped published human iPS-derived skin organoid scRNA-seq data[89] to our
489 HDCA (**Methods**). We selected human skin organoid as this organoid closely mimics prenatal skin structural

490 composition[80] and has a proposed anatomical origin (craniofacial region derived from neural crest
491 differentiation)[89]. We observed that skin organoid fibroblasts largely mapped to 3 major mesenchymal
492 groups: progenitors, skin, and skeletal (**Fig. 3k-l; Fig. S20a**). The skeletal population is consistent with a
493 known off-target cartilage population of human skin organoid. Within the progenitor group we identified 4
494 *ZIC*^{hi} populations (*ZIC2*$^{+}$*PCSK9*$^{+}$*_scalp*, *ZIC2*$^{+}$*ANGPT*$^{+}$*_primary_meninx*, *ZIC2*$^{+}$*WNT4*$^{+}$*_primary_meninx*,
495 *ZIC2*$^{+}$*FIBIN*$^{+}$*_spine)* (**Fig. 3l-m**). We had earlier identified these *ZIC*^{hi} progenitor subtypes in the embryo
496 head and spine region (**Fig. 3d, i**), and *ZIC* proteins are linked to neural crest induction as well as meningeal
497 fibroblast precursor identity[90,91]. We also identified *RUNX2*$^{+}$ progenitors, linked to the development of
498 skeletal fibroblast populations, and *SATB1*$^{+}$ progenitors. These organoid progenitor populations expressed
499 similar gene signatures to the equivalent human embryo populations, with the exception of *SATB1*$^{+}$
500 progenitors (**Fig. S20b)**. The presence of our newly defined progenitors within the human skin organoid
501 model provides further validation of these populations consistent with their anatomic location in human
502 embryos. Finally, the cellular census powered by the HDCA allows resolution of fine cell subtypes within
503 diverse *in vitro* systems. For instance, within the off-target cartilage population, which previously has been
504 broadly labelled as *POSTN*$^{+}$ fibroblasts only, we identified *GDF5*-expressing *Interzone cells*, which
505 characterize the initial mesenchymal condensation of synovial joints and give rise to chondrocytes[22], as well
506 as *Chondrocytes* themselves, *Tenocytes*, *Skeletal myocytes*, *Osteoblasts*, and *Perichondrium* (**Fig. S20c-d**),
507 illustrating the value of our resource to decode unrecognised stromal heterogeneity in organoid models.
508 Overall, we illustrate how our spatially-resolved profiles of developmental human stroma, defined across
509 millions of cells, can be used to define novel complexity in fibroblast progenitors in human organoid models
510 and resolve fine heterogeneity through reference-query mapping. Our HDCA atlas and its associated web
511 portal can thus be used as a powerful resource for interpreting future organoid and organ-specific datasets
512 as well as guiding organoid differentiation.
513
514 ***Organ-specific blood endothelial specialisation in development***
515 The second distributed system we explored was the nascent vascular system. The vascular endothelium
516 consists of a layer of endothelial cells (EC) lining all vessels, exhibiting remarkable molecular and functional
517 diversity and adapting to the specific demands and microenvironments of distinct tissues and organs[92,93].
518 Single cell atlases have been used to profile molecular identity across the adult human vasculature, revealing

519 the transcriptional state of arteries and veins in addition to organ-specialised capillary programs
520 underpinning endothelial function and phenotype[94–96]. Studies have also revealed that EC molecular
521 heterogeneity emerges during embryogenesis and is shaped by local tissue cues in mice[97,98]. Developmental
522 EC data in humans however is restricted to individual fetal organ studies[99–101]. Here, we leveraged the HDCA
523 to isolate and analyse 94,681 ECs and provide the first systematic, multi-organ, and temporally resolved
524 atlas of EC states across human gestation.
525
526 Unsupervised clustering of the HDCA ECs revealed 37 transcriptionally distinct cell types (**Fig. 4a**). To
527 annotate these subsets, we curated canonical marker genes from an adult human vascular atlas[94], and cross-
528 referenced them with transcriptional signatures and structured metadata from the public datasets within the
529 HDCA (**Fig. S21a**). All major endothelial compartments were represented, including arterial, venous,
530 capillary, and lymphatic ECs (**Fig. S21b**). Consistent with adult profiling[94], arterial and venous ECs largely
531 formed organ-agnostic clusters (**Fig. 4b**), with central nervous system and spleen harbouring distinct
532 arteriolar populations from the canonical common arterial EC cluster shared between organs. Alongside a
533 large capillary EC cluster shared by multiple organs (**Fig. S21c**), certain organs, including the heart, lung,
534 central nervous system, and liver, possessed transcriptionally distinct capillary EC subsets, reflecting their
535 fully differentiated adult counterparts[94].
536
537 To more precisely resolve the timing of EC specialisation, we computed organ-specific gene programs (GPs)
538 using differentially expressed genes from scRNA-seq data of adult ECs across multiple organs[94]. For each
539 developing organ, we generated a subset of the ECs and computed GP scores corresponding to their adult
540 counterparts, before plotting these GPs across developmental time using generalised additive models (GAM)
541 (**Fig. 4c**). In brain, heart and liver, organs with transcriptionally distinct capillary ECs, adult-like gene
542 programs were already present as early as 6PCW. In order to validate this finding, we integrated our
543 suspension data with 163,973 ECs sampled across 11 Xenium sections of the 6PCW human embryo (**Fig.
544 S21d**), and performed label transfer from our HDCA EC atlas to Xenium data (**Fig. 4d**). We additionally
545 detected the expression of canonical organotypic capillary-enriched genes within the Xenium data, including
546 the blood brain barrier endothelial markers *SLC7A5*$^{+}$, which was expressed in ECs lining the maturing
547 CNS[102], ECs within the endocardium expressing the endocardial marker *POSTN*$^{+}$[103], and ECs within the

Figure 4: Spatiotemporal analysis of blood vascular and lymphatic endothelial heterogeneity during human development

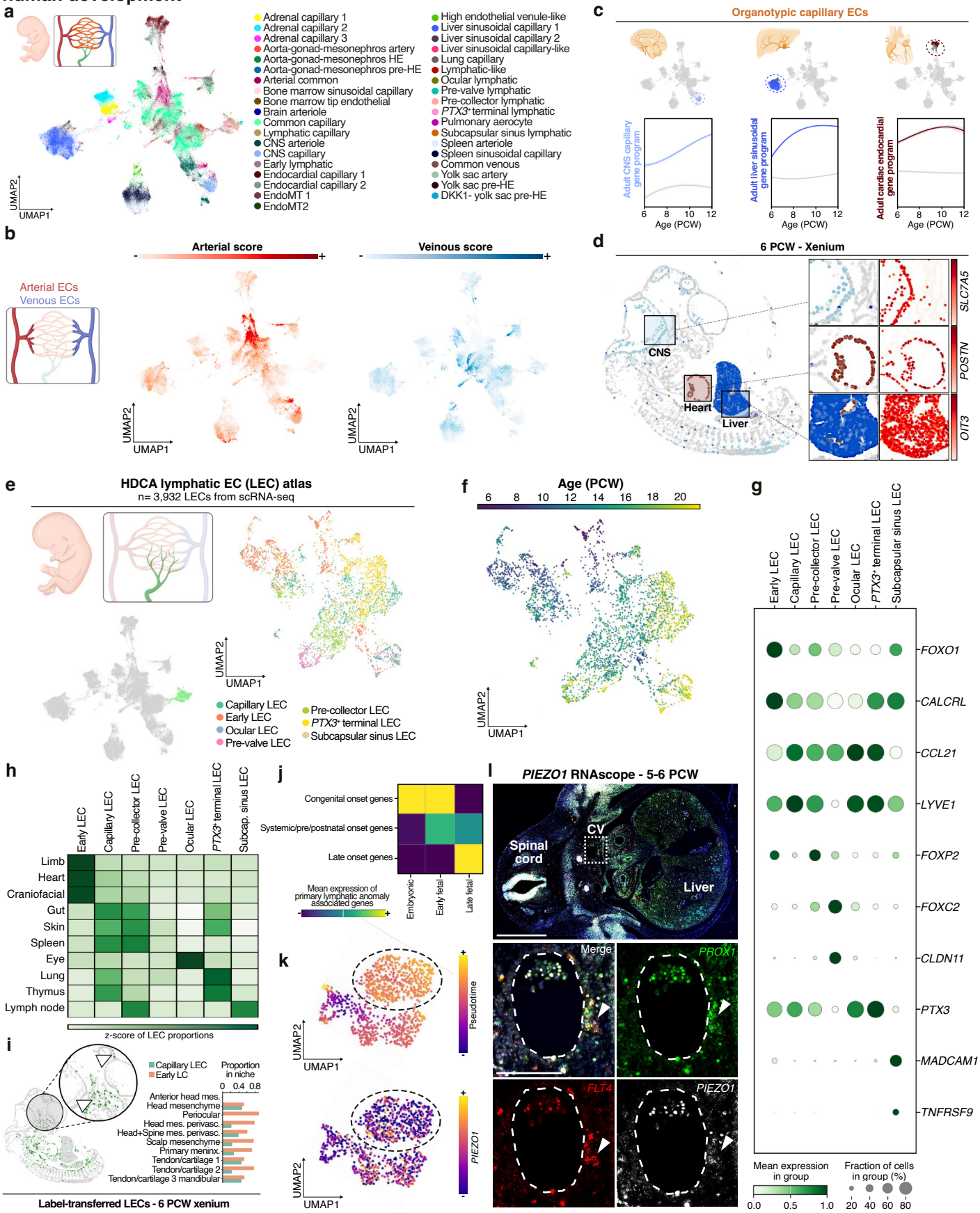


Figure 4: Spatiotemporal analysis of blood vascular and lymphatic endothelial heterogeneity during human development

a) UMAP visualisation of HDCA endothelial cells (EC, n=94,681 cells) generated using scRNA-seq, resolving 37 transcriptionally distinct EC types. **b)** Feature plots showing arterial (above, red) and venous (below, blue) gene signatures on the EC UMAP. **c)** Generative additive model (GAM) analysis of organ-specific capillary subsets in the heart (left), liver (middle) and central nervous system (CNS, right), with above UMAPs depicting endocardial cells, liver sinusoidal endothelial cells and CNS capillaries, respectively. Adult gene programs sourced from previously published data were plotted against time in these EC populations (coloured lines) and plotted against randomly sampled ECs from the atlas (grey lines). The faded edge of each line represents the 99% confidence interval. **d)** ECs were isolated from Xenium data, and label transfer and analysis of marker genes was performed. Label transfer from scRNA-seq to Xenium lead to successful labelling of ECs in the CNS (top), heart (middle) and liver (bottom) with corresponding organ-specific capillary-enriched marker genes shown to the right of each subpanel. Section S133 is shown and is representative of three sections analysed. **e)** UMAP visualisation of HDCA lymphatic ECs (LEC, n=3,932 cells), subclustered to resolve 7 transcriptionally distinct LEC types. **f)** UMAP of LECs are plotted here by developmental stage. **g)** Dotplot showing canonical marker genes for each LEC subtype. **h)** Matrix plot with the colour representing z scores for percentage of LEC subtypes across organs. **i)** Label transfer from scRNA-seq to Xenium demonstrates lymphatics to appear in the cranial region from multiple serial sections of a human embryo at ~6PCW, shown in section S133 representative of three sections analysed. **j)** Calculation of scores for genes associated with primary lymphatic anomalies with congenital onset, systematic / pre- or postnatal onset and late onset, with scores shown over embryonic, early fetal and late fetal stages within LECs of the HDCA. **k)** Pseudotime analysis of early LECs (top, n=995 cells) and featureplot showing *PIEZO1* expression across the trajectory (bottom). Dotted circle indicates early pseudotime. **l)** RNAScope for *PIEZO1* of axial sections of a 5-6 PCW human embryo, alongside the lymphatic markers *PROX1* and *VEGFR3*. The top panel shows low magnification, with the bottom panels showing the boxed region magnified. The white arrowhead shows the co-expression of *PIEZO1, PROX1* and *VEGFR3* in cells within the wall of and adjacent to the cardinal vein (CV). Scale bars 200µm.

548 embryonic liver which were enriched for the liver sinusoidal endothelial marker, *OIT3*[104]. In contrast, shared
549 capillary ECs exhibited diffuse and organ-agnostic distribution (**Fig. S21e**). These results suggest that organ-
550 specific endothelial transcriptional programs are established during early organogenesis, and that molecular
551 hallmarks of specialised adult ECs are detectable during the first trimester.
552
553 ***Lymphatic endothelial heterogeneity during human development***
554 We next characterised the lymphatic EC (LEC) compartment, identifying a distinct rare cluster of 3,932
555 LECs within the HDCA, identified by their expression of canonical lymphatic endothelial markers including
556 *PROX1, FLT4/VEGFR3,* and *PDPN* (**Fig. S22a**). The lymphatic system in adults is highly morphologically
557 specialised[105–108], and emerging evidence shows LECs possess organ-specific structural, molecular and
558 functional features[109–111]. While scRNAseq has revealed insights into lymphatic development in model
559 organisms[112,113], studies in human development have been restricted to individual organs[114–116], and a
560 systematic analysis of the profile and timing of LEC specialisation during human development is required.
561 Unsupervised clustering of LECs revealed 7 transcriptionally distinct subsets (**Fig. 4e**). One cluster was
562 enriched in cells from early embryonic stages (6–10PCW) (**Fig. 4f**), which we termed early LECs. Marker
563 gene analysis showed early LECs expressed *FOXO1* and *CALCRL* (**Fig. 4g**), genes essential for the
564 regulation of lymphatic development in mouse models[117–121]. Conversely, the remaining 6 clusters,
565 containing cells from later developmental stages, showed transcriptional similarities with known LEC
566 subsets in adulthood, including *LYVE1*$^{+}$*CCL21*$^{+}$ capillary[122,123], *FOXP2*$^{+}$ pre-collecting[124] and
567 *FOXC2*$^{+}$*CLDN11*$^{+}$ pre-valve[125,126] LECs, ocular LECs[127], *PTX3*$^{+}$ terminal lymphatics[128,129] and
568 *MADCAM1*$^{+}$*TNFRSF9*$^{+}$ LECs associated with the lymph node subcapsular sinus[109,130]. To investigate organ-
569 specific patterns, we quantified LEC subtype abundance across tissues (**Fig. 4h**). Capillary and pre-collecting
570 LECs were broadly distributed, while ocular and subcapsular sinus LECs were spatially restricted to the eye
571 or lymph node, respectively. *PTX3*$^{+}$ terminal LECs, known to interact with innate immune cell subsets[128,129],
572 were enriched in thymus and mucosal or barrier organs, including gut, skin, and lung. Intriguingly, early
573 LECs were most abundant in the limb buds, heart, and craniofacial regions. Whereas heart lymphatic
574 development has been profiled in mouse studies[131], and cardiac progenitors with potential for lymphatic
575 differentiation have been identified using lineage tracing experiments[132–134], the existence of LECs within
576 the human developing cranium contrasts with prior mouse literature. These studies suggest that lymphatics

577 in this region, such as those of the meninges and ocular regions, emerge postnatally in mouse[135–137]. We
578 corroborated these findings with our Xenium data, within which label transferred LECs were detected within
579 the head region (**Fig. 4i**). Leveraging niche annotations to more carefully resolve their anatomical location,
580 capillary LECs and early LECs were detectable in 9 out of 10 head niches, including the primary meninx,
581 the primitive layer of mesenchyme giving rise to the meninges, where central nervous system lymphatics
582 have been reported in human adults[138–142].
583
584 ***Spatiotemporal expression of genes associated with primary lymphatic disorders***
585 When genes essential for lymphatic development and function are disrupted, the result is a spectrum of
586 disorders leading to the build-up of interstitial fluid within the tissue extracellular space[143]. Collectively,
587 these are known as primary lymphatic anomalies (PLA)[144]. To assess the relevance of our analysis of human
588 developing LECs to PLA, we profiled expression of nine genes implicated in PLAs, commonly used in
589 clinical variant screening panels (**Fig. S22b; Supplementary Table 14**). Gene expression was stratified by
590 developmental age and grouped into embryonic, early fetal and late fetal stages, with mean expression scores
591 calculated for genes associated with congenital, pre- or postnatal, and late-onset (> 1 year after birth) forms
592 of PLA (**Fig. 4j**)[144]. Genes associated with congenital and early onset PLA were preferentially enriched at
593 embryonic stages, whereas genes linked to late onset PLA were enriched at fetal stages. This temporal
594 stratification is concordant with clinical observations[143], and suggests that early-onset disorders arise from
595 variants in genes associated with disruption of lymphatic specification, whereas late-onset disease-associated
596 genes may act during later stages of lymphatic maturation.
597
598 One gene of particular interest was PIEZO1, which is associated with generalised lymphatic dysplasia, an
599 early onset PLA[145,146]. PIEZO1 encodes a mechanosensitive ion channel, and although biallelic loss of
600 function variants in *PIEZO1* have been associated with fetal hydrops[145,146], its also implicated
601 in postnatal lymphatic valve formation and maturation, a later phenomenon and currently the most studied
602 role of PIEZO1 in the lymphatic system[147,148]. To investigate this further, we examined PIEZO1
603 expression in *FOXO1*$^{+}$ *CALCRL*$^{+}$ early LECs. Pseudotime analysis predicted *PIEZO1* expression even
604 at the earliest stages of the trajectory for these early LECs (**Fig. 4k**). As vasculature development is poorly
605 recapitulated in human organoid models[80], we corroborated

606 our findings in model organisms used to profile lymphatic development. We examined recently published
607 single cell data of LECs and associated ECs at the onset of lymphatic emergence in mice[112] and zebrafish[113]
608 embryos (**Fig. S22c-f**). Using the original author cell type annotations, we found that ECs committed to
609 lymphatic endothelial fate were enriched in the expression of the murine and zebrafish orthologs of *PIEZO1*,
610 alongside *FOXO1* and *CALCRL*, markers used to identify early LECs within the HDCA. To validate this
611 finding within a human context, we performed RNAscope in 5–6PCW embryos, demonstrating expression
612 of *PIEZO1* in cells co-expressing the pan-LEC markers, *PROX1* and *FLT4/VEGFR3*. These cells were
613 localised within and immediately adjacent to the cardinal vein (**Fig. 4l**), a major embryonic site of LEC
614 emergence in mammals and zebrafish[108].
615
616 Thus, by leveraging the HDCA, we have defined the timing and molecular heterogeneity underpinning
617 human endothelial diversification during development. We reveal that organ-specific capillary EC
618 transcriptional programs, particularly within the brain, heart and liver, are established as early as the first
619 trimester of human development. Resolving lymphatic endothelial transcriptomes across time and space
620 demonstrated the early establishment of molecular heterogeneity, with disease-associated lymphatic genes
621 such as *PIEZO1* expressed at stages that precede the acquisition of fully differentiated lymphatic cell states.
622
623 ***Temporal reconstruction of the developing human peripheral nervous system***
624 The third and final distributed system we explored was the peripheral nervous system (PNS). The
625 development of the human PNS, encompassing both sensory and autonomic domains, remains poorly
626 characterised. This is largely due to its anatomical complexity, which spans the entire organism, and its dual
627 developmental origin from two transient multipotent populations arising at the neural plate border: cranial
628 placodes and neural crest cells (NCCs). Although somatic motor neurons contribute to peripheral nerve
629 circuitry, they are neural tube–derived and therefore developmentally CNS in origin, and are not considered
630 further here. Placode- and NCC-derived states have been characterised largely in model organisms[149,150],
631 while human studies have focused on isolated PNS compartments[151,152], limiting our understanding of
632 lineage contributions and tissue organisation.
633

634 Our single cell and spatial atlas identified the full repertoire of NCC-derivatives (*SOX10*$^{+}$ glial and *PRPH*$^{+}$
635 neuronal) across the embryo and placode-derived sensory neurons in cranial regions (*SIX1, EYA1, PRPH*)
636 (**Fig. 5a; S23a-d**). NCCs are predicted to give rise to sensory, autonomic, and mesenchymal lineages, with
637 a pseudotemporal analysis in mice suggesting that sensory specification precedes autonomic and
638 mesenchymal differentiation[149]. In contrast, within the broad NCC *(SOX10, FOXD3)* compartment we
639 observed sensory *(PRDM12)*, autonomic *(PHOX2B)*, and mesenchymal (*PDGFRA*) populations co-existing
640 at 6PCW (**Fig. 5a-b**), precluding inference of a clear temporal ordering between sensory and autonomic
641 specification. To identify the most undifferentiated NCC-derived cell states and characterise lineage
642 differentiation, we applied Cell2fate[153] as it outperforms existing RNA velocity models in reconstructing
643 known cell fate trajectories and produces interpretable RNA velocity gene modules. This analysis confirmed
644 that sensory and autonomic populations co-exist as transcriptionally active, early lineage-committed states,
645 from which downstream sensory and autonomic trajectories are predicted to emerge (**Fig. S23e**). Multipotent
646 sensory NCCs were predicted to progress through sensory neuronal progenitor states to form sensory neurons
647 (*CNTN2*), while autonomic NCCs were predicted to develop into autonomic neuronal progenitor states
648 (*PHOX2B, ASCL1)* and diverse autonomic neuronal derivatives (enteric (*RET*), sympathetic (*DBH, TH,*
649 *MAOA*), parasympathetic (*SLC5A7*)), as well as sympathoblasts (*STMN2*) and chromaffin cells (*CHGA*)
650 (**Fig. 5a; Fig. S23e**). These data provide the first comprehensive delineation of human PNS cell types and
651 their lineage relationships, as summarised in **Fig. 5c**.
652
653 NCCs are classically described as a progenitor population that is molecularly distinct from schwann cell
654 precursors (SCPs), with each population defined by partially non-overlapping marker sets. However, recent
655 studies in mice have proposed the existence of a multipotent NCC/SCP 'hub' state characterised by co-
656 expression of NCC and SCP markers[154]. Consistent with this model, both sensory and autonomic NCC
657 populations (*SOX10, FOXD3*) in the HDCA data universally co-expressed SCP markers (*MPZ*, *PLP1*) from
658 the earliest stages examined (**Fig. 5d**). Despite the uniform co-expression of NCC and SCP markers, it was
659 still possible that an NCC-only state exists transiently and is difficult to capture in human embryos. We thus
660 leveraged the iPS-derived skin organoid system in which NCC-like populations are induced and can be
661 interrogated from the earliest stages of differentiation[80,89]. The skin organoid dataset[89] (**Fig. S23f-g**) similarly
662 corroborated early co-expression of NCC and SCP markers, with no clearly separable NCC-only state

Figure 5: Temporal reconstruction of the developing human peripheral nervous system

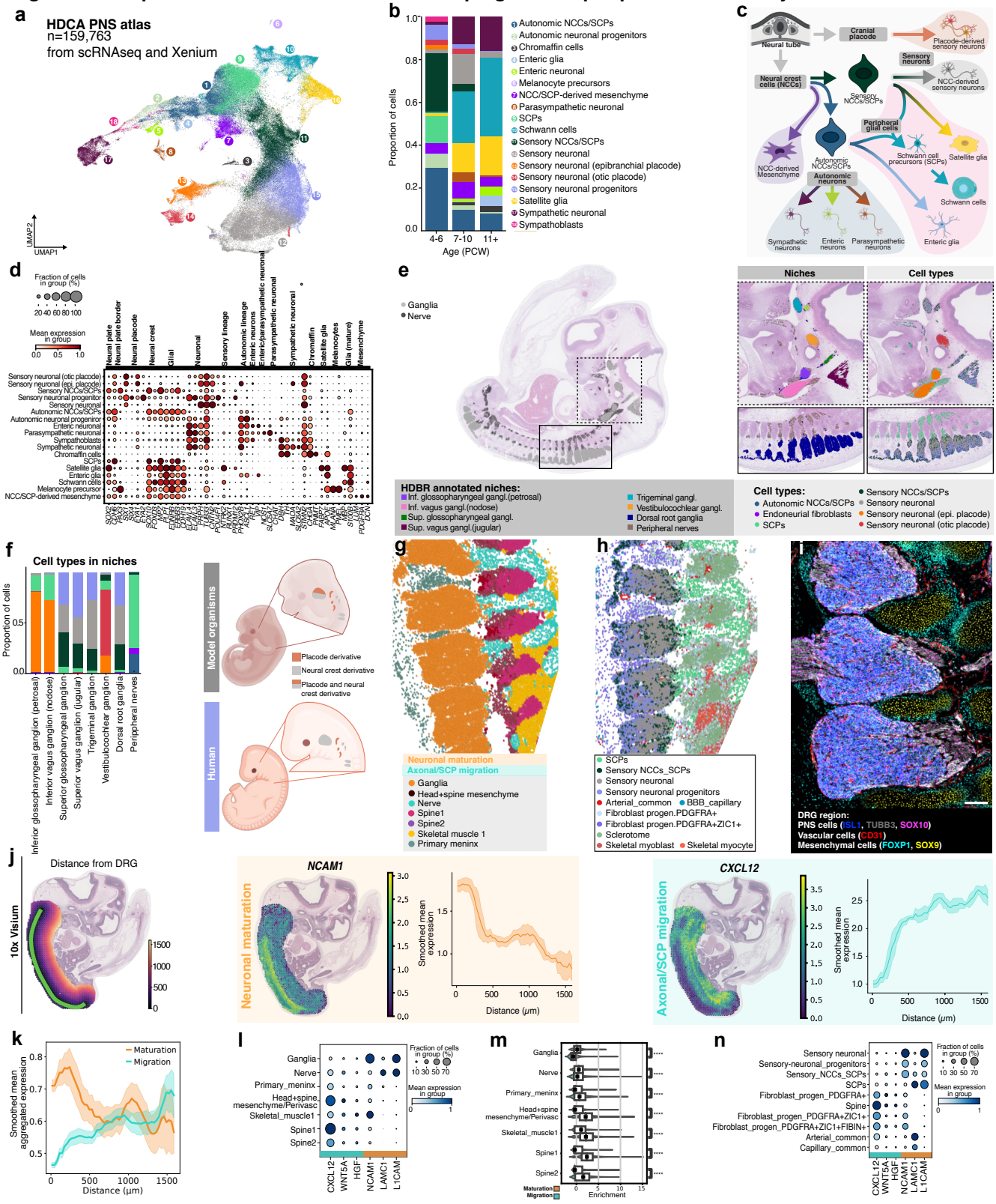


Figure 5: Temporal reconstruction of the developing human peripheral nervous system

a) UMAP visualisation of the PNS (n=159,763 cells) from both scRNA-seq and Xenium. Age, technologies, and lineage enrichment are shown in **Fig. S23a**. **b)** Barplot depicting cell type proportions per age. **c)** Schematic illustrating PNS trajectories. The diagram highlights the developmental and spatial progression of PNS lineages, and the emergence and differentiation of sensory and autonomic neuronal populations alongside associated progenitor and glial trajectories. Arrows indicate lineage relationships and transitional states, providing an overview of how distinct PNS cell types arise. **d)** Dotplot of marker gene expression for PNS populations. **e)** Left: Spatial location of the niches covering the PNS overlaid on H&E of a 6PCW embryo (section S133), and a UMAP coloured by cell type (3 slides). Right: Spatial location of the PNS populations of an embryo at 6PCW (left). Zoom into specific PNS areas (right) into head ganglia region (top) and DRG (bottom). H&E, HDBR annotated niches and cell types. **f)** Barplots depicting the proportion of PNS cell types found in each region, and schematic of cell origin per niche compared to model organisms. **g-h)** Xenium 10x niches (g) and cell types (h) in the DRG region of a 6PCW embryo (section S133) highlighting: neuronal maturation (left) and axonal/SCP migration (right). **i)** Section of human embryo (7PCW) showing DRG region immunostained for vascular (CD31), PNS (ISL1, TUBB3, SOX10) and mesenchymal cells (FOXP1, SOX9). Scale bar 200µm. **j)** Expression of *CXCL12* and *NCAM1* by distance from the DRG (proximal to distal planes) using Visium (section_121). Colour scale overlaid on H&E representative images indicate the normalised gene expression. Line plots depict the gene expression by distance from the DRG. **k)** Line plots of the mean gene expression by distance of genes associated with neuronal maturation and axonal/SCP migration (see **S21c** for individual ligands). **l)** Dotplot with the gene expression of ligands associated with axonal/SCP migration and neuronal maturation in niches in the DRG region of a 6PCW embryo (Xenium section S133). **m)** Violin plots depicting the enrichment of the gene signatures for neuronal maturation and axonal/SCP migration for the niches found in the DRG region of a 6PCW embryo (Xenium section S133). **n)** Dotplot of gene expression of ligands associated with axonal/SCP migration and neuronal maturation in the cell types within the DRG of 6PCW embryo (Xenium section S133). Dot size represents the proportion of cells within each group expressing a given gene. Color intensity indicates the scaled mean expression level. Gene expression values were standardised across genes to facilitate comparison. Statistical significance was determined by two-sided, Wilcoxon rank sum test with Bonferroni multiple testing correction. p < ****: p <= 1.00e-04.

663 detected, suggesting that such a state is either extremely transient or that prenatal human NCCs broadly adopt
664 an SCP-like transcriptional identity.
665
666 The population of NCC-derived mesenchyme in our dataset co-expressed canonical NCC (S*OX10, FOXD3*)
667 and mesenchymal markers (*PDGFRA, LUM, DCN*) (**Fig. 5a, c-d**). Notably, this population expressed
668 *TWIST1*, a transcription factor shown to be critical for enabling trunk NCCs, which are primarily neuralgenic
669 with limited mesenchymal potential compared to cranial NCCs[155], to acquire ectomesenchymal potential in
670 mice[149]. We also detected *FRZB* and *ZIC* gene expression (**Fig. S23h**), consistent with the above described
671 *ZIC*^{hi} mesenchymal populations (**Fig. 3d, i**), which are also characterised by high *TWIST1* expression but no
672 longer express NCC markers (**Fig. S23h**). The presence of these markers within the NCC-derived
673 mesenchyme supports the above interpretation that *ZIC*^{hi} mesenchymal lineages are indeed NCC-derived
674 and suggests that the NCC-derived mesenchyme population captured here represents a progenitor state
675 giving rise to these cells. Consistent with this model, we detected an early mesenchymal population in the
676 skin organoid NCC[89] that co-expressed an NCC transcriptional signature alongside mesenchymal markers
677 *FRZB*, *TWIST1*, and *ZIC*s at d6 (**Fig. S23g**) before the emergence of the 3 *ZIC*^{hi} populations at d26 (**Fig. 3l**).
678 Notably, we found that the organoid NCCs, which possess a cranial identity[80,89], gave rise to a greater
679 abundance of mesenchymal cells relative to neural derivatives over time (**Fig. S23i**), consistent with the
680 enhanced mesenchymal potential of cranial NCCs *in vivo*[155]. Together, these findings support a conserved
681 NCC-to-ectomesenchyme transition across *in vivo* and organoid systems.
682
683 The NCC compartment in our dataset also encompassed a diverse range of peripheral glia (Schwann cells,
684 satellite and enteric glia) and melanocytes (**Fig. 5a-d**). Temporal analysis revealed that neuronal
685 specification occurs around 6PCW, while glial cells (**Fig. 5a-b, Fig. S23a**), as well as a glial maturation
686 signature (**Fig. S23j**), emerged at later developmental stages. Thus consistent with model organisms[154,156]
687 and as previously shown for human DRG development[151], overall, human peripheral nervous system
688 development exhibits a temporal hierarchy, with neurogenesis preceding gliogenesis.
689
690
691

692 ***Neural crest- and placode-derived sensory neurons***

693 Sensory ganglia have been shown to arise from placodal and neural crest lineages in model
694 organisms[149,150,157,158]; however, it is unknown whether the same developmental origins apply in humans.
695 Transcription factors associated with placode- and neural crest–derived sensory lineages have been well
696 characterised in model organisms and are frequently used as lineage markers; however, the extent to which
697 these markers reliably distinguish placode- versus NCC-derived populations during human development
698 remains unclear. We found that placode-associated transcription factors *EYA1/2* and *SIX1/4*[157,158] were not
699 restricted to placode-derived populations in human development, but were transiently expressed in NCC-
700 derived sensory neuron progenitors, consistent with reports from human dorsal root ganglion (DRG)
701 studies[151,159] (**Fig. 5d**). In contrast, *SOX10* and *FOXD3* specifically marked NCC-derived populations, while
702 *PRDM12* identified the sensory NCC-derived lineage, enabling clear discrimination between placode- and
703 NCC-derived sensory neurons (**Fig. 5d**). These distinctions were further supported by 6PCW single nucleus
704 multiome analysis showing that placode- and NCC-derived cells clustered separately (**Fig. S23k**), with
705 chromatin accessibility at placode-associated genes broadly shared, whereas accessibility at NCC-associated
706 genes, including *PRDM12*, was specific to NCC-derived sensory neurons (**Fig. S23l).**

707

708 We next mapped these populations spatially within PNS niches identified by NicheCompass (**Fig. 2**) across
709 three Xenium tissue sections in which PNS niches were abundant (**Fig. 5e**). Ganglia niches corresponded to
710 sensory ganglia at 6PCW and were enriched for neuronal transcripts (**Fig. 5e**). Of note, we only detected
711 sensory ganglia, not parasympathetic, sympathetic or enteric ganglia, suggesting that although autonomic
712 and sensory NCC/SCPs co-exist (**Fig. 5a, Fig. S23a**), sensory neuronal differentiation and ganglia
713 maturation precedes autonomic. Leveraging these spatial and transcriptional data, we showed evidence for
714 the developmental origins of human sensory ganglia for the first time (**Fig. 5f**). Consistent with animal
715 studies[160,161], we identified both placode-derived (vestibulocochlear, nodose and petrosal) and NCC-derived
716 (dorsal root, jugular and superior glossopharyngeal) sensory ganglia in humans, each with distinct molecular
717 signatures (**Fig. S23m**). Notably, we found that the human trigeminal ganglion consisted purely of NCC-
718 derivatives, in contrast to the mixed origins reported in model organisms[162]. Together, these analyses provide
719 the first human-specific map of sensory ganglia origins, defining placodal and neural crest derivation,
720 revealing key differences from model organisms.

721 ***The PNS development is guided by tissue niche interactions***

722 Following ganglia formation, PNS development proceeds through axonal outgrowth toward target tissues,
723 accompanied by the migration of SCPs along extending axons, where they support axon guidance and
724 subsequent glial differentiation. To identify those axonal structures, we examined the second PNS cell–rich
725 niche identified by NicheCompass across 6PCW Xenium embryos, which was distinct from ganglia. We
726 termed this niche *Nerve* (**Fig. 2**), as it spatially mapped to peripheral nerve structures on H&E sections, both
727 within the spinal region and along axonal extensions emerging from cranial sensory ganglia, and was
728 enriched for NCC/SCP and SCP populations (**Fig. 5e**). In addition to axons, the *Nerve* niche also identified
729 PNS compartments not discernible by H&E, such as the inferior hypogastric plexus (**Fig. S24a**),
730 underscoring the power of our niche approach to resolve distinct gene expression profiles within
731 histologically featureless tissues and across systems (beyond the pharyngeal endoderm shown in **Fig. 2**).
732 NCC-derived mesenchymal populations also localised along axons, as well as around ganglia and in regions
733 overlapping with *ZIC*$^{+}$ mesenchymal populations in the primary meninx, spine/DRG and scalp regions, again
734 suggesting a direct lineage-relationship (**Fig. S24b**).

735

736 Using the NCC/SCP-rich *Nerve* niche as a proxy for tissue innervation, we found that, at 6PCW, PNS cells
737 have not yet colonised organ parenchyma, in contrast to the mesenchymal and vascular compartments, which
738 we have shown already display organ-specific identities at this stage (**Figs. 3-4**), indicating that
739 mesenchymal and vascular niches are established prior to PNS innervation. The only exception was the gut,
740 where autonomic NCC/SCP populations were detected that start forming the enteric nervous system (**Fig.**
741 **S24a**), even earlier than previously reported[163]. In line with their later developmental emergence, mature
742 enteric neurons[163], enteric glia[163], Schwann cells[151,159], satellite glia[151,159], and chromaffin cells[152] were not
743 observed in the 6PCW Xenium data (**Fig. 5b**).

744

745 To investigate how tissue niche interactions shape PNS development, we focused on the DRG and spinal
746 nerve region, where sensory neurons undergo maturation and axonal outgrowth within a region composed
747 of spatially adjacent PNS, mesenchymal and vascular niches and cells (**Fig. 5g-i**). We analysed spatially
748 resolved ligand–receptor signalling across the previously identified niches within this region (**Fig. 5i-j**) using
749 a literature- and Gene Ontology-curated set of signalling pathways implicated in neuronal maturation and

750 axonal/SCP migration[164–166]. We first mapped ligand expression onto a tissue section using Visium to
751 examine their spatial distribution (**Fig. 5j**). This analysis revealed opposing dorsoventral patterns: ligands
752 associated with SCP migration and axon-associated guidance, such as CXCL12[167], showed a ventrally
753 enriched, gradual increase consistent with gradient-based, chemoattractive signalling, whereas neuronal
754 maturation ligands, including *NCAM1*[168], exhibited an inverse dorsally enriched distribution consistent with
755 locally acting, short-range signalling (**Fig. 5j-k; Fig. S24c**). To identify the niche and cellular sources of
756 these ligands, we analysed their expression across both niches and constituent cell types (**Fig. 5l-n**).
757 Consistent with sensory neuron maturation occurring within ganglia, maturation-associated ligands
758 (including *NCAM1*, *LAMC1* and *NTF3*; **Fig. 5l**) were enriched in the ganglia niche (**Fig. 5m**) and PNS-
759 derived populations (**Fig. 5n**), although expression was not restricted to this compartment (**Fig. 5l)**. In
760 contrast, cues associated with axonal outgrowth and SCP migration (e.g. *CXCL12*, *WNT5A* and *HGF*)
761 showed higher relative expression within the Nerve niche and adjacent mesenchymal and vascular niches
762 surrounding extending axons (**Fig. 5l-n**). Analysis of distance-aware cell–cell interaction and downstream
763 signaling activity further confirmed that maturation-associated ligand–receptor pairs scored highest at short
764 intercellular distances, supporting a local signalling mode, whereas migration-associated interactions were
765 less distance constrained (**Fig. S24d-e**). These findings reveal spatially organised niche signalling
766 programmes that coordinate PNS migration, differentiation and integration into embryonic tissues.

767 Notably, we detected expression of key regulators of sensory ganglia formation within the NGF signalling
768 pathway, with *NGF* and *NTF3* expressed by mesenchymal cells and the receptor *NTRK1* expressed by
769 sensory neurons (**Fig. S24f**). Given that genetic loss of these components in mice results in hypoplastic or
770 absent DRG, extensive sensory neuron apoptosis, and impaired peripheral axon outgrowth[169–171], these data
771 suggest that NGF signalling pathways implicated in congenital peripheral neuropathies are active at this
772 developmental stage. Together, our integrated atlas defines conserved and human-specific principles of PNS
773 lineage specification, spatial organisation and intersystem communication during prenatal development.

774 **Discussion**

775 This study presents the Human Developmental Cell Atlas (HDCA, n=186, k>4.6 million cells/nuclei, 4–
776 22PCW), a unified resource integrating community-generated, organ-focused single cell/nucleus
777 developmental datasets, alongside our complementary single cell and spatial transcriptomics (Xenium,
778 Visium) data of 6PCW human embryos to fill a critical gap in late embryogenesis. Our HDCA data serves
779 as a resource to contextualise human cells in development and disease, and those recapitulated within *in vitro*
780 organoid systems. We've ensured the HDCA's future reusability and continued biological insights by
781 making data and models (scVI, CellTypist, scArches, NicheCompass) available, and creating a bespoke
782 web portal for equitable access (https://cellatlas.io/studies/hdca/; private until publication).
783

784 The HDCA provides a temporally and spatially resolved view of how the human body is built. Using
785 NicheCompass, we provide a comprehensive decomposition of the human embryo into 114 distinct tissue
786 niches sampled across 11 sagittal whole embryo Xenium sections, representing a step-change in how cell
787 types have been assessed and contextualised in human developmental single cell atlases to date. Niche
788 decomposition allowed us to identify potential drivers of tissue patterning, and to define histologically
789 featureless niches with distinct gene expression profiles, such as within the *Ant_head_mesenchyme* niche,
790 which would not have been possible with cell types and spatial data alone. We make our NicheCompass
791 models publicly available, enabling future community embryo niche identification both without expert
792 histopathological annotation and even within a small number of sections.
793

794 We show how fibroblast progenitors diversify early to establish tissue niche architecture. Whether these
795 tissue-specific embryonic fibroblast progenitors mature *in situ*, or migrate, remains an important open
796 question. The HDCA also allowed us to resolve complexity within rare mesothelial populations (epithelial
797 monolayers surrounding the outside of visceral organs) and the LPM populations from which they are
798 derived. We identify shared and unique gene expression patterns between mesothelial cells across the whole
799 embryo body, and detail, for example, LPM subtypes with distinct transcriptomic and spatial profiles.
800 Detailing mesothelial ontogeny in development has implications for our understanding of aberrant programs
801 in disease, such as mesotheliomas. Such detailed interrogation of mesothelium and fibroblasts have only
802 been made possible through the pooling of these respectively scarce and abundant cells across datasets.

803 Further, our integrated analysis across organs and systems enabled us to delineate NCC-derived
804 mesenchymal populations, in human and across *in vivo* and organoid models.
805
806 Secondly, within the vascular system, we provide a comprehensive prenatal, cross-tissue atlas of human
807 endothelial cells, encompassing arterial, venous, capillary and lymphatic compartments across development.
808 Our analyses revealed that endothelial diversification occurs early, preceding the establishment of fully
809 differentiated and functional organ systems[172]. We demonstrate early and rapid diversification of specialised
810 vascular beds within the developing heart, central nervous system and liver, and that organ-specific capillary
811 transcriptional programs[94,95] are detectable as early as the mid-first trimester. This timing is in contrast to
812 recent scRNA-seq studies of murine vascular development, in which organ-specific EC identity has been
813 reported to emerge mid-gestation in mouse embryos[173]. The early emergence of these adult-like capillary
814 programs in humans suggest that endothelial specialisation is coupled to early tissue patterning[98], or intrinsic
815 cues[174,175], rather than arising as an adaptation to organ function in later development. Similarly, lymphatic
816 endothelial heterogeneity was established early, including a distinct subset of early lymphatic cells in
817 humans, enriched for genes with evolutionarily conserved roles in lymphatic development such as *FOXO1*,
818 *CALCRL* and *PIEZO1*[117–121,145,147,148]. Notably, transcriptional signatures corresponding to adult lymphatic
819 subtypes, including *PTX3*-expressing terminal capillary lymphatics[128,129], were already detectable during
820 human fetal development, indicating that lymphatic diversification precedes the acquisition of mature
821 lymphatic patterning and organ-specific molecular phenotypes[111]. In contrast to mouse models, in which
822 cranial lymphatics emerge postnatally[135–137], we detect evidence for lymphatic cells in the human
823 craniofacial region as early as 6PCW. Altogether, these findings support a model in which organ-specific
824 capillaries and lymphatic endothelial heterogeneity are specified early in human gestation, providing a
825 molecular substrate upon which later vascular remodelling, structural specialisation and functional
826 maturation are built. These findings also have clinical implications, as we illustrate that the developmental
827 timing of lymphatic gene expression has concordance with the timing of onset of primary lymphatic
828 anomalies[143].
829
830 Finally, within the PNS, we present a spatially resolved framework of human PNS ontogeny, capturing a
831 broad repertoire of likely placode- and neural crest cell (NCC)-derived states, including sensory and

832 autonomic neurons, peripheral glia, NCC-derived mesenchyme, sympathoblasts, chromaffin cells and
833 melanocytes. We resolved lineage contributions across distinct ganglia niches and directly assessed placodal
834 versus NCC origins of human sensory ganglia *in situ*. Overall, the inferred developmental origins of human
835 sensory ganglia were largely consistent with those of model organisms[160,161], with one notable exception: in
836 contrast to the mixed placodal and NCC contributions reported in animal models[162], the human trigeminal
837 ganglion was composed exclusively of NCC-derived sensory neurons. Although we cannot formally exclude
838 a transient placodal contribution at stages earlier than those sampled, the absence of a placodal signature in
839 the human trigeminal ganglion (despite clear and distinguishable placode-derived populations in the petrosal,
840 nodose and vestibulocochlear ganglia) argues against a substantial placodal input. This raises the possibility
841 that, unlike in model organisms[162], placodal input is functionally dispensable for trigeminal sensory roles in
842 humans, or that NCC-derived sensory neurons can functionally substitute by adopting molecular and
843 functional programs typically associated with placodal lineages, noting that lineage-specific roles within the
844 trigeminal ganglion remain poorly defined even in model organisms[157,158,176] and motivating future lineage-
845 tracing and functional studies. Beyond lineage specification, our spatial analyses indicate that human PNS
846 development involves temporally coordinated interactions between neighbouring tissue niches. Analysis of
847 spatially resolved, known ligand-receptor interactions[164–166] revealed a spatial division of labour in which
848 mesenchymal and vascular compartments, established prior to PNS innervation, form an early scaffold that
849 supports axonal outgrowth and Schwann cell precursor migration. Together, these findings underscore the
850 ability of an holistic atlas such as the HDCA to reveal intersystem cross-talk events required for niche and
851 tissue development.
852
853 The HDCA spans tissues and distributed tissue networks, and can therefore act as a springboard for
854 researchers with diverse interests to explore prenatal cell types of interest. Decades of model organism
855 developmental biology research provided crucial literature to contextualise many of the cell types and
856 lineages in this study, however, these were never detailed in humans before, or at such scale. We report both
857 human-specific aspects of development and those which are conserved with model organisms, with the hope
858 that this study can more widely act as a human developmental cell reference for understanding
859 developmental disorders in model systems. Organ biologists can use the HCDA to understand how cellular
860 composition and tissue architecture forms across time and space, and researchers interested in inter-system

861 cross-talk can leverage this unique resource to comprehensively evaluate such interactions at a niche-level,
862 as here in the context of the PNS-mesenchyme. Our data and models provide a foundation for future research
863 integrating new developmental datasets, we hope even bridging the gap to the earliest implantation stages.
864
865 Looking toward the utility of a future HDCAv2[5], important challenges to tackle include data generation from
866 undersampled tissues, such as from the endocrine system and the time points between implantation and
867 4PCW. Spatial profiling at sequential timepoints, e.g., 8PCW and 10PCW (but especially those from less
868 represented embryonic and late fetal timepoints) combined with niche inference, would be a powerful
869 approach to capture transient tissue microenvironments, especially given that our fixed sampling at only
870 6PCW was able to resolve embryonic tissue drivers, granular niche capture and cellular composition of
871 pharyngeal niches and hitherto undescribed cranial niches. While we provide the most comprehensive
872 assessment of niche capture to date from 11 sagittal sections and >4 million cells, more comprehensive
873 profiling is required to fully capture niche composition across the whole embryo. More widely, inferring
874 molecular information from H&E is an open problem in computational and clinical research[177], and niche
875 decomposition provides one method for tackling this open problem at organism-scale by discretisation into
876 functionally relevant histological groupings. The translational applications of the HDCA are wide-ranging,
877 particularly with our Cherita webportal integration with the DDG2P dataset
878 (https://www.ebi.ac.uk/gene2phenotype)[43] allowing for both gene-cell and gene-niche investigation into
879 developmental disorders for which genetic variants have been identified.
880
881 In summary, this study represents a major milestone toward a unified and complete HDCA that will be a
882 transformative resource for research and clinical communities. Together, by enabling development to be
883 analysed through spatially and temporally defined tissue niches, our HDCA shows that, similar to
884 architecture, form follows function as tissue structure emerges from coordinated interactions between cells,
885 tissues and systems.

886 **Methods**

887 The Human Developmental Cell Atlas (HDCA) was constructed by integrating 2 complementary sources of
888 single cell RNA-sequencing data: (i) a whole-embryo (WE) scRNA-seq dataset generated de novo in this
889 study, and (ii) a large collection of publicly available community datasets (Community data). The
890 annotations for the community datasets were informed by the original author-provided labels and
891 subsequently harmonised across studies. In contrast, the WE dataset lacked pre-existing cell type annotations
892 and extensively manually curated through literature informed gene markers and expert input, augmented by
893 machine learning models generated using the community data. The final HDCA was generated by jointly
894 integrating and reconciling information from both datasets, leveraging the strengths of de novo annotation
895 in the WE data and the diversity of community-derived annotations. In this study we specifically focussed
896 on: mesenchyme, endothelial cells and the peripheral nervous system. Below, we describe the strategies for
897 separate sample preprocessing, quality control, integration, clustering and cell type annotation of WE and
898 Community data and how they informed annotation of HDCA. All code supporting the methods and analyses
899 presented in this study will be made available at https://github.com/haniffalab/hdca-paper upon publication.
900

901 ***Community data: Public human developmental single cell data acquisition and preprocessing***

902 We collected 21 community scRNAseq datasets, comprising 18 publicly available ones and 3 obtained
903 through internal communications, representing diverse tissues across the entire body of the human embryo
904 and/or fetus (**Supplementary Table 1, 5, 7**). The selected datasets were accessed in 2022, and fulfilled the
905 following 3 criteria: 1) consistent technology (10x Genomics Chromium), 2) minimum 750 cells post-QC
906 per organ (see below for QC thresholds), 3) the majority of annotation labels provided were granular, i.e., to
907 the cell type level 'monocyte' rather than lineage 'myeloid'. To render comparability between pre-existing,
908 pre-processed single-cell RNA-sequencing (scRNA-seq) studies, together comprising over 3.6 billion cells,
909 we implemented a scalable unification pipeline that systematically harmonised the feature space,
910 standardised metadata, and concatenated the expression matrices into a single analytic object. For each
911 dataset, the raw count matrix was downloaded along with the accompanying cell meta information. After
912 removing trivial cell types annotated by the original studies (for example, "Doublets"), we used a loose
913 quality control criterion by requiring a minimum number of expressed genes per cell > 200, and obtained a
914 total of 3,652,263 cells representing all major cell populations in the human embryo (**Supplementary Table**

915 **1**). Thirty-one curated metadata fields, including biological attributes (organ, age), donor-specific identifiers
916 (integration_donor), and technical descriptors such as sequencing_platform, were harmonised and appended
917 to the cell-level annotation table, thereby establishing a consistent observational schema. Comprehensive
918 provenance information, including dataset identifiers, original publication details, accession codes, and
919 acquisition metadata, is catalogued in **Supplementary Table 5**. For datasets obtained through
920 restricted-access agreements or pre-publication data-sharing arrangements, relevant ethical approvals and
921 consent frameworks are documented in the original source publications.
922
923 ***Community data: Realignment and gene space harmonisation***
924 We used the preexisting data from the original publications. To ensure all datasets were aligned to GRCh38,
925 Miller et al 2020[178], Sridhar et al 2020[179], and Yu et al 2021[180] were remapped to GRCh38 2020-A as outlined
926 in **Supplementary Table 1**. To resolve assay- and release-specific disparities in gene nomenclature for
927 external datasets, we employed the pre-release IDTrack algorithm[181] (see
928 https://github.com/theislab/idtrack), a graph-based identifier harmonization framework designed to reconcile
929 discrepancies in gene nomenclature across large-scale transcriptomic datasets. Variability in gene identifiers
930 arises not only from the use of distinct annotation systems such as Ensembl and RefSeq, but also from
931 differences between annotation releases, which evolve over time and contribute to inconsistencies in gene
932 symbol conventions. In our integration effort, datasets derived from various genome builds and annotation
933 versions were standardized by mapping all gene identifiers to a unified target space defined by HUGO Gene
934 Nomenclature Committee (HGNC) gene symbols corresponding to Ensembl release 107. Gene symbols that
935 mapped unambiguously to a single Ensembl identifier were retained without modification, whereas
936 one-to-many relations were adjudicated by prioritising biologically verified target identifiers, and genes
937 exhibiting irreconcilable or failed conversions were removed. Dataset-level statistics, such as the numbers
938 of genes whose mapping changed only because of one-to-many conflicts or because of alternative, equally
939 valid one-to-one targets, were logged to enable traceability, and the full conversion table was stored within
940 the unified object for subsequent audit. When combined with WE scRNA-seq data to form the complete
941 HDCA data, only genes which intersected with the genes post IDTrack were retained.
942
943

944 ***Community data: Highly variable genes selection pipeline***

945 To enable robust, scalable, and biologically meaningful downstream analyses across millions of single cells,
946 we developed a comprehensive feature selection pipeline designed to eliminate confounding technical
947 signals, enforce rigorous quality control, and identify highly variable genes (HVGs) in a batch-aware
948 manner. Beginning with a consolidated expression matrix comprising 3,611,521 single-cell transcriptomes
949 profiled across 57,009 genes, we removed 516 genes annotated to the cell-cycle process, 400
950 immunoglobulin heavy- and light-chain genes encoding B-cell receptors, and 227 αβ or γδ T-cell receptor
951 genes, whereas mitochondrial and cytoplasmic ribosomal transcripts were retained to preserve markers of
952 cellular fitness and translational capacity; this operation reduced the data matrix to 3,609,265 cells by 55,866
953 genes. Subsequent quality control removed 2,256 cells (0.06 % of the total) with fewer than 200 detected
954 genes, and filtered 11,490 genes supported by fewer than 5 UMI counts across the entire cohort, yielding a
955 high-fidelity dataset of 3,609,265 cells and 44,376 genes that served as the input for variable-gene selection
956 step.

957

958 Raw gene counts were normalised using *sc.pp.normalise_total* function with target sum as 10^4 and then a
959 shifted log transformation was applied using the *sc.pp.log1p* function. The batch_key was defined by the
960 Cartesian product of 4 covariates —donor identity, biological unit, sample status, and sequencing platform,
961 resulting in 192 batches. For each gene, mean expression was constrained to the interval $0.05 \leq \mu \leq 6$ to
962 exclude extremely rare or constitutively expressed features, after which 219 genes falling outside the
963 specified expression bounds were discarded. Normalized dispersion values were computed independently
964 for each batch and also aggregated globally. From the global rankings, the top 12,288 genes ($2^{14} - 2^{12}$)
965 exhibiting the highest dispersion were retained as candidate HVGs for downstream analysis. Concordance
966 analysis revealed a median Spearman correlation between global and batch-specific dispersion rankings that
967 exceeded that expected by chance, and the intersection between the top 10,000 HVGs derived from the global
968 model and the per-batch selections reached 75.4 % across batches, attesting to the stability of the procedure.
969 Moreover, a Wilcoxon rank-sum test comparing dispersion profiles obtained with and without batch
970 correction yielded $P < 10^{-15}$, underscoring the statistical benefit of the batch-aware strategy.

971

The final unified dataset, comprising 57,009 harmonised genes and 3,611,521 quality-controlled cells contains (i) gene-level annotations that include harmonised ENSEMBL names and intersection flags, (ii) complete provenance records listing the unification_id, preprocessing_id, IDTrack parameters, and per-study conversion logs in adata.uns, and (iii) the full set of 31 standardised metadata columns. By enforcing stringent quality and formatting standards, this uniformization strategy ensures that all downstream analyses operate on a reproducible and harmonized data foundation, thereby enabling rigorous and unbiased inference in single cell transcriptomic studies.

Community data: Integration with scVI

To address the multifaceted challenge of integrating single cell RNA sequencing data spanning multiple organs, developmental time points, and assays, we implemented a cell-type-agnostic training and validation strategy based on the single cell Variational Inference (scVI)[182] framework. An extensive hyperparameter search was conducted in 2 phases to identify a suitable network architecture and training scheme for large-scale data integration as described below. This approach was designed to accommodate the inherently high-dimensional and sparse nature of early developmental atlases (i.e., genes × cells × time × organ) and to mitigate the numerous technical confounders frequently encountered in large-scale integration, including unbalanced sampling across organs and time points, protocol-specific batch effects (e.g., fresh versus frozen, nuclei versus whole cell, 5′ versus 3′ sequencing). In addition to scVI, we evaluated other cell-type-agnostic methods, including scPoli (deployed without prototype initialization), TarDis (using batch labels only as the disentanglement target and did not incorporate any auxiliary covariates like cell-type annotations), Scanorama, Harmony, and principal components analysis (PCA), each subjected to a 40-run hyperparameter tuning process.

In the first phase of model training, an extensive hyperparameter search spanning more than 4 thousand configurations of scVI was conducted to identify a suitable network architecture and training scheme for large-scale data integration. Model hyperparameters, encompassing the dimensionality of hidden layers, the timing and magnitude of Kullback–Leibler regularization warm-up, dropout rates, and learning rates, as well as dataset-level considerations such as the strategy for selecting highly variable genes (union versus intersection across datasets) and the number of input features were systematically varied. For each

1001 configuration, multiple batch definitions were tested to ascertain whether biological attributes or technical
1002 factors (e.g., donor identity, sequencing platform) should be explicitly encoded as batch variables, reflecting
1003 the heterogeneity intrinsic to multi-organ, multi-time-point studies. A preliminary benchmarking pipeline
1004 computed R^2 scores on differentially expressed genes, along with a rapid subset of integration quality metrics
1005 from the scIB suite, thereby enabling high-throughput evaluation across many candidate models. Early
1006 convergence patterns revealed pronounced sensitivity to parameters controlling regularisation strength—
1007 particularly weight decay and the duration of KL warm-up, and to architectural depth, whereas broader
1008 sampling of activation functions and optimiser variants yielded diminishing returns. Accordingly, search
1009 density was adaptively shifted toward the high-impact subspace, with occasional manual refinements applied
1010 when intermediate checkpoints indicated latent spaces with superior topology preservation but suboptimal
1011 reconstruction. Empirical validation of these early trials led to a narrowed set of hyperparameters that
1012 balanced model expressiveness and training stability while mitigating overfitting.
1013
1014 To further refine integration performance, an additional 40 training configurations of scVI were also
1015 explored to refine the integration pipeline and corroborate the hyperparameter settings derived from the
1016 initial large-scale optimization. We also evaluated other cell-type-agnostic methods, scPoli, TarDis,
1017 Scanorama, Harmony, and principal components analysis (PCA), following the 40-run hyperparameter
1018 tuning process. Importantly, the knowledge gained during the preliminary 4,000-configuration scVI sweep
1019 guided parameter selection for subsequent runs of autoencoder based models (scVI, scPoli[183] and TarDis[184]),
1020 thereby aligning each model's architecture and training strategy with empirically validated best practices
1021 while retaining sufficient variation to identify the most robust approach. Scanorama and Harmony were
1022 executed with their default parameters. Benchmarking of the models were performed using scIB and
1023 scTRAM[45] metrics to quantify the balance between batch removal, biological conservation, and trajectory
1024 representation fidelity.
1025
1026 To establish a baseline for evaluating integration performance, we computed differentially expressed genes
1027 (DEGs) using the Wilcoxon rank-sum test implemented in scanpy's method, `tl.rank_genes_groups`. Genes
1028 were ranked by log-fold changes, adjusted p-values (padj<0.05), and significance thresholds, with results

1029 stored in the AnnData object's unstructured (uns) metadata. This preliminary analysis enabled post hoc
1030 validation of cell-type-specific markers preserved during integration.
1031
1032 A cell-type-agnostic modelling paradigm was adopted for several reasons rooted in the intrinsic complexity
1033 and partial annotation of large-scale developmental atlases. Because emerging cell populations in early
1034 embryonic or fetal stages may lack definitive markers, strictly label-based strategies risk propagating bias
1035 from uncertain or incomplete annotations. Furthermore, dynamic cell definitions can shift abruptly across
1036 time points or organs, rendering static category assignments unreliable. By avoiding explicit cell-type priors,
1037 we aimed to preserve flexibility in uncovering rare or transitional states, as well as accommodate divergent
1038 nomenclature across datasets without prematurely imposing a fixed annotation framework. This approach
1039 thereby emphasized the inherent structure of the data over potentially erroneous or heterogeneous cell-type
1040 labels.
1041
1042 ***Community data: Benchmarking integration and trajectories with scIB and scTRAM***
1043 To systematically evaluate how effectively different integration methods and model configurations preserve
1044 both batch correction and developmental biology in the multi-organ human developmental cell atlas, we
1045 implemented a dual-framework benchmarking strategy combining the established scIB[44] protocol with
1046 scTRAM[45], a trajectory-fidelity metric suite specifically designed to assess preservation of developmental
1047 progressions in integrated representations. Cell embeddings with different hyperparameters were obtained
1048 using scVI, scPoli, TarDis, Harmony, Scanorama, and principal components analysis (PCA). The scVI,
1049 scPoli, and TarDis models were trained with architectures and hyperparameters derived from the extensive
1050 4,000-configuration optimization described in the *Data Integration* section, ensuring each autoencoder-
1051 based method operated under empirically validated best practices. Harmony and Scanorama were executed
1052 with their default parameters as specified in their respective implementations. For all methods, the batch_key
1053 was defined as mentioned above in the selection of highly variable genes section.
1054
1055 *scIB: Conventional integration quality assessment*
1056 The scIB benchmarking framework quantifies integration performance through a comprehensive suite of
1057 complementary metrics conceptually partitioned into 2 categories: biological conservation and batch

removal. Biological conservation metrics assess the extent to which intrinsic cellular structure and lineage relationships are maintained following integration, encompassing measures such as cell-type average silhouette width, normalized mutual information, adjusted rand index, isolated label F1 score, and local inverse Simpson's index computed with respect to cell-type labels (cLISI). These metrics collectively evaluate whether the integrated representation preserves discrete cell-type identities, maintains neighbourhood composition, and prevents unwarranted amalgamation of biologically distinct populations. Batch removal metrics quantify the effectiveness with which technical covariates are suppressed, including graph connectivity, batch average silhouette width, k-nearest-neighbour batch effect test, principal component regression, and local inverse Simpson's index computed with respect to batch labels (iLISI). These measures assess whether batch-associated structure has been adequately mitigated without introducing artificial discontinuities or over-correction artefacts.

For each integration method, individual metrics were computed on latent representations using hierarchical cell-type annotations at both broad cell level and fine-grained subtypes as ground-truth references. Metric values were min–max scaled within their respective categories, directionalities harmonized such that higher values uniformly indicate superior performance, and subsequently averaged within biological conservation and batch removal groups before computing a composite aggregate score. This normalization procedure enabled direct ranking of integration strategies according to the joint criterion of biological fidelity and technical robustness, thereby facilitating systematic comparison across methods with heterogeneous performance profiles.

scTRAM: Trajectory-specific fidelity evaluation

While scIB effectively quantifies batch correction and cluster alignment, it does not capture whether embeddings preserve the differential-geometric properties of continuous cellular transitions that underpin lineage commitment and differentiation dynamics. To address this limitation, we applied scTRAM, single cell trajectory representation metrics, a principled benchmarking framework that systematically evaluates how well integrated representations retain ground-truth developmental trajectories across complementary failure modes.

1087 Developmental trajectories were formalized as directed acyclic graphs G_{ref} = (V, E), where vertices V
1088 correspond to discrete cell types and directed edges E represent validated state transitions between cell types.
1089 We curated 3 exemplar reference systems with distinct developmental architectures: hematopoiesis (57 cell
1090 types, 58 directed connections), germline development (8 cell types, 7 connections), and the peripheral
1091 nervous system (12 cell types, 11 connections). These reference graphs captured diverse trajectory topologies
1092 ranging from hierarchical branching structures in hematopoiesis to linear progressions in germline
1093 development, ensuring comprehensive evaluation across multiple scales of developmental complexity.
1094 Ground-truth trajectories and corresponding cell-type annotations for each system are presented in
1095 **Supplementary Figure 7**.
1096
1097 scTRAM evaluates trajectory fidelity by comparing inferred developmental structures in the latent
1098 representation against the curated reference lineage graph G_{ref} through 3 complementary metric groups that
1099 decompose trajectory preservation into distinct performance axes. Topology conservation metrics
1100 (adjacency-based) assess whether the branching architecture and connectivity patterns of the reference graph
1101 are maintained in the integrated representation, quantifying structural similarity through measures including
1102 graph edit distance, spectral distance, and topological data analysis approaches, among others. Cell-order
1103 continuity metrics (pseudotime-based) evaluate whether cells are arranged in the correct developmental
1104 progression along directed edges, testing temporal monotonicity through rank correlations such as
1105 Spearman's ρ, dynamic time warping distance, and concordance indices, alongside additional ordering-based
1106 measures. Trajectory dominance metrics (embedding-based) determine whether the developmental manifold
1107 remains the primary source of variation in the latent space rather than being obscured by batch effects or
1108 other nuisance structure, employing manifold preservation measures such as neighbourhood conservation,
1109 trustworthiness, and geodesic stress, supplemented by directionality and continuity assessments. For each
1110 integration method, diffusion-based pseudotime and cell-type connectivity graphs were inferred from the
1111 latent representation, then compared against reference pseudotimes and adjacencies derived from G_{ref} using
1112 the full battery of metrics within each group. Individual metric scores were subsequently aggregated within
1113 their respective categories to produce topology conservation, cell-order continuity, and trajectory dominance
1114 summary scores for systematic comparison across integration methods.
1115

Edge-resolved and lineage-stratified analysis

To enable fine-grained performance assessment across developmental transitions, we partitioned each reference lineage graph into constituent trajectory segments using an automated decomposition algorithm. This algorithm identified hierarchical functional modules by extracting induced subgraphs over all branch points and their descendants, enumerated branch-to-leaf paths from the trajectory root to all terminal cell types, traced linear chains from each bifurcation point's immediate successors until subsequent branch points, and computed the longest path through the developmental hierarchy via dynamic programming. Candidate subgraphs were retained only if they contained at least 3 nodes, exhibited weak connectivity, and possessed exactly one root node. This multi-resolution decomposition yielded hundreds of trajectory segments spanning diverse developmental transitions from stem cell activation to terminal differentiation, ensuring balanced coverage of distinct developmental pathways and enabling localized performance analysis that aggregate metrics would obscure.

To localize performance differences along developmental paths, we computed per-edge scTRAM scores across the 3 curated lineage graphs. For each directed edge in G_{ref}, we identified all trajectory segments containing that edge and calculated the mean performance rank across all applicable metrics within those segments, thereby deriving edge-specific fidelity scores for each integration method. Differences between methods were assessed using two-sided paired comparisons after multiple-testing correction via the Benjamini–Hochberg procedure at a 5% false discovery rate threshold. This edge-resolved analysis revealed non-uniform, stage-specific behaviour wherein fidelity varied continuously along developmental progressions and models selectively preserved or distorted particular developmental segments.

In parallel, trajectory-specific analysis was performed by grouping the decomposed segments according to major lineage classes (myeloid, lymphoid, megakaryocyte-erythroid-mast, and stem/progenitor), computing mean ranks within each group, and evaluating whether methods exhibited systematic biases in handling different branching complexities or developmental contexts. Performance was additionally stratified by scTRAM metric group (pseudotime-based, adjacency-based, embedding-trajectory) to assess whether models optimized for preserving topological connectivity sacrificed fidelity in pseudotemporal ordering or vice versa.

1145 *Integrated benchmarking framework*

1146 For each latent representation, scIB and scTRAM metrics were computed independently and subsequently
1147 combined to provide a comprehensive evaluation of integration quality. Individual scTRAM metrics were
1148 min–max scaled and averaged within their respective groups (topology conservation, cell-order continuity,
1149 trajectory dominance) before computing a trajectory representation aggregate score. Integration methods
1150 were ranked according to both conventional scIB criteria (biological conservation and batch removal) and
1151 trajectory-specific scTRAM criteria, with particular attention directed toward identifying trade-offs between
1152 batch correction and lineage preservation. Qualitative visual inspections of lineage-restricted subsets
1153 complemented quantitative assessments, ascertaining preservation of rare transitional states and detecting
1154 unwarranted amalgamation of biologically distinct populations. Particular scrutiny was directed toward the
1155 retention of developmental trajectories across organs and time points and toward the attenuation of artefacts
1156 attributable to protocol heterogeneity, most notably differences between fresh and frozen preparations,
1157 nuclei- versus whole cell captures, and 5′- versus 3′-biased library chemistries.

1158

1159 This dual-framework approach enabled rigorous model selection that balanced conventional integration
1160 objectives with trajectory-specific requirements, ensuring that the chosen representation not only achieved
1161 effective batch correction and cluster purity but also faithfully preserved the continuous manifold structures
1162 essential for downstream analyses of developmental progression, lineage commitment, and transcriptional
1163 dynamics along differentiation paths.

1164

1165 ***Community data: Cell type annotation***

1166 In the first iteration, we adopted an automated method for annotating the community data based on the author
1167 provided annotations for each constituent dataset, followed by manual curation for standardising cell type
1168 names. We trained a logistic regression-based model for each previously published dataset using CellTypist[30]
1169 (version 1.4.0), publicly available at *https://www.celltypist.org/models*, with the function *celltypist.train*
1170 (tag-value pairs: max_iter=1000, feature_selection=True, top_genes=300). We cross-predicted cell identities
1171 for each pair of datasets using these models with the function *celltypist.annotate*, and based on these
1172 predictions we standardised cell type names across datasets to achieve a common naming scheme.
1173 Specifically, we collapsed cell types from different datasets only if they could be reciprocally predicted and

1174 thus corresponded to a single cell type (e.g., alignment between "*Multiciliated precursor*" from Miller et al.
1175 2020[178] and "*Deuterosomal*" from He et al. 2022[20]). Moreover, we located and subsequently refined a coarse
1176 cell type only if it was divisible into subtypes from another dataset and the latter could be uniquely mapped
1177 back (e.g., subdivision of "*pro_B*" from Garcia-Alonso et al. 2022[185] into "*PRO_B*" and "*LARGE_PRE_B*"
1178 from Suo et al. 2022[24]). The resulting cell types underwent another round of manual curation, including
1179 inspection of marker gene expression and renaming of cell types, which resulted in a final list of 536 cell
1180 types curated from 766 author-defined original cell annotations
1181 (*https://www.celltypist.org/encyclopedia/Human_Whole_Embryo/v1*).
1182
1183 To test the performance of this model, we split the entire data set into a set of training (90%) and the
1184 remaining set of test dataset (10%). A CellTypist model was then trained from the training set and used to
1185 predict cell types for the test dataset. Specifically, we defined one iteration as one mini-batch training, with
1186 each batch containing 1,000 cells randomly selected from the training dataset. A CellTypist model was
1187 derived after each iteration and was then used to predict cell types for the test dataset. This process was
1188 repeated over 30 epochs, with each epoch consisting of 100 mini-batches; in other words, a total of 3,000
1189 iterations were performed (**Fig. S1e**). The model performance was assessed for each cell type separately
1190 using 3 metrics: precision ("*sklearn.metrics.precision_score*"), recall ("*sklearn.metrics.recall_score*"), and
1191 F1 score ("*sklearn.metrics.f1_score*"). Lastly, we averaged these scores across cell types to obtain the
1192 performance curves, showing the model performance as a function of training iterations (**Fig. S1e**). Second,
1193 we tested the performance of the final model (i.e., the model trained from 100% of the data) on an
1194 independent dataset profiling the human embryo using sciRNA-seq3 (Cao et al. 2020)[47], and visualised the
1195 cell type correspondence between the model and the tested dataset using *celltypist.dotplot* (**Fig. S1f**). In the
1196 second iteration the resultant cell types were manually crosschecked with original author annotations per
1197 leiden cluster (resolution=3), considering organ and lineage diversity to annotate the cells for the community
1198 data. This model is used to increase the power of manual annotations of the de novo Whole Embryo
1199 dataset. In the second iteration the resultant cell types were manually crosschecked with original author
1200 annotations per leiden cluster (resolution=3), considering organ and lineage diversity to annotate the
1201 cells for the community data.
1202

1203 ***Whole embryo: sample acquisition and processing***

1204 Human embryos for scRNAseq/multiome (samples F137, F147, F158) and for spatial Xenium, Visium,
1205 RareCyte (samples SOB25, SOB26) were collected by the MRC– Wellcome Trust-funded Human
1206 Developmental Biology Resource (HDBR; http://www.hdbr.org; MR/R006237/1, MR/X008304/1 and
1207 226202/Z/22/Z) with written consent and approval from the Newcastle and North Tyneside NHS Health
1208 Authority Joint Ethics Committee (23/NE/0135) (**Supplementary Table 1**). One human embryo for spatial
1209 MACSIMA (sample SM74P7) was provided by the HuDeCA Biobank (Inserm) following French ethical
1210 and legal guidelines. Donors provided written informed consent prior to sample procurement. Ethical
1211 approval for the use of these materials was granted by the Agence de la Biomédecine (Saint-Denis La Plaine,
1212 France; Authorization N° PFS19-012) and the INSERM Institutional Review Board (IRB00003888).

1213

1214 ***Whole embryo: scRNAseq sample processing, library preparation and sequencing***

1215 Each of the three whole human embryo samples for scRNAseq/multiome were dissected into twelve
1216 anatomical regions using a scalpel (**Supplementary Table 2; Fig. S1a-e**). Unfortunately, Embryo 1 (Haniffa
1217 Lab sample ID F137) was received with one lower limb missing. Each tissue region was stored in
1218 HypoThermosol® Preservation Media (STEMCELL Technologies) for a maximum of sixteen hours
1219 refrigerated. Each tissue was then washed with phosphate buffered saline (PBS, Sigma-Aldrich) before
1220 mincing with a scalpel. The minced tissue was digested with 1.6 mg/ml type IV collagenase (Worthington)
1221 in RPMI (Sigma-Aldrich) at 37°C for 30 min with intermittent agitation. Digested tissue was then passed
1222 through a 100µm cell strainer and 0.25% (w/v) trypsin (Sigma-Aldrich) was added to any undigested on the
1223 strainer mesh for 5 min before adding RPMI supplemented with 10% (v/v) heat inactivated fetal bovine
1224 serum (FBS), 1% (v/v) penicillin streptomycin and 1% (v/v) L-glutamine. Cells were then collected by
1225 centrifugation (500 g for 5 min at 4°C) and manually counted using an Improved Neubauer hemocytometer.
1226 Cells from each tissue region were stained with DAPI (Sigma-Aldrich) at a final concentration of 3µM and
1227 passed through a 35µm filter (Falcon) before FAC-sorting to select for only live, single cells (**Fig. S1c**).

1228

1229 Immediately after sorting, cells were then manually counted using an Improved Neubauer hemocytometer.
1230 Where possible, 4 channels of cells from each tissue region were loaded onto the 10x Genomics Chromium
1231 Controller to target 10,000 cells per channel. Embryo 1 (Haniffa Lab sample ID F137) was processed with

1232 the 10x Genomics Chromium Single Cell 5' Reagent Kit v1 and embryos 2 and 3 (Haniffa Lab sample ID
1233 F147 and F158) were processed with the 10x Genomics Chromium Next-GEM Single Cell 5' Reagent Kit
1234 v2, as per the manufacturer's instructions. Sequencing libraries were also processed with the 10x Genomics
1235 proprietary reagents. Libraries were then sequenced using Illumina NovaSeq S4 Flow Cells to target a
1236 minimum sequencing depth of 50,000 UMIs per cell.
1237
1238 ***Whole embryo: scRNAseq alignment and quality control***
1239 Whole embryo (WE) scRNAseq data was mapped to the human reference genome (GRCh38-2020-A) using
1240 CellRanger (v6.0.1). Empty drops were excluded from analysis through use of filtered CellRanger matrices.
1241 2 low quality lanes were excluded: WS_wEMB10202360 had low fraction of valid barcodes;
1242 WS_wEMB11031943 had low number of cells detected. This produced an initial count matrix of size
1243 1,183,081 cells x 36,601 genes).
1244
1245 Filtering was performed to select for cells expressing >1,000 UMIs and >=500 genes, and genes expressed
1246 in >=5 cells (QC step 1/2), leaving a count matrix of size 1,138,964 cells x 34,584 genes. Low quality cells
1247 as determined by high mitochondrial content were calculated and flagged for later evaluation alongside
1248 annotation information. We aimed to use the strictest threshold available while ensuring the majority of any
1249 annotated cell type category was not excluded from downstream analysis. Cells with mitochondrial reads
1250 <5%, 10%, 15% and 20% were evaluated for selection, and cells with <10% mitochondrial reads fulfilled
1251 criteria above and were selected for downstream analysis. To assess doublet contribution to the scRNAseq
1252 data, Scrublet (v0.2.3) was applied to each lane independently, obtaining per cell doublet scores. An
1253 exclusion threshold of >(median + (3 * median absolute deviation)) Scrublet score was applied on a per cell
1254 basis, as previously described[16,17].
1255
1256 To investigate the potential of maternal contamination we followed a similar protocol outlined in Dann and
1257 Suo et al[24]. Every 10x lane from each sample was put through Souporcell (v2.4) with genotype clusters
1258 k=1/2/3 models to represent the likelihood of single or multiple genotypes (2nd genotype represents
1259 maternal, 3rd represents control and artefacts). Optimal model selection was determined using BIC (bayesian
1260 information criterion). Most 10x lanes were determined to be a singular genotype. The optimal model for

1261 lane WS_wEMB10202336 was k=3, but given no technical replicates were available for this sample:section
1262 (F137 torso section 1) these data were retained. Of note, 2 genotypes were detected for lanes
1263 WS_wEMB10202351, WS_wEMB10202353 and WS_wEMB10202356 (from sample F137 torso sections
1264 5 and 6). The cells contributing to the smaller genotype cluster were flagged as potential maternal cells and
1265 recorded in metadata prior to annotation, then removed after annotation was complete.
1266
1267 Following removal of cells based on high mitochondrial content, doublet status, and maternal contaminant
1268 status (QC step 2/2), the count matrix was of size 1,029,326 cells x 34,584 genes.
1269
1270 To investigate the level of ambient RNA within the scRNAseq data, Cellbender (v0.2.0) was run for each
1271 high quality lane. The resultant soup-corrected CellBender filtered matrix summed to 1,363,423 cells x
1272 36,601 genes. A 'soup' matrix was then generated by taking the difference between CellBender-soup-
1273 corrected counts and the non-corrected CellRanger counts. Given the more permissive cell calling employed
1274 by CellBender vs CellRanger, we had to access the equivalent counts for the non-corrected 1,363,423 cells
1275 through use of unfiltered CellRanger matrices. CellBender was run on a per sample basis with the following
1276 parameters: expected yield of 10,000 cells per lane, total-droplets-included=20,000, fpr=0.01, epochs=150.
1277 Parameters were adapted by lane based on CellRanger barcode rank plot: WS_wEMB10202336 expected
1278 cells parameter was lowered to 5000; WS_wEMB11031950 total-droplets-included was set to 60,000;
1279 WS_wEMB12142148 total-droplets-included was set to 35,000; WS_wEMB12142131 failed due to
1280 insufficient background, therefore no corrected raw counts were obtained. These soup matrices then
1281 facilitated investigations into ambient RNA contribution across each sample. The top 200 gene counts from
1282 each sample's soup matrix were recorded (**Supplementary Table 4**) and the top 50 were recorded in the data
1283 for reference in downstream annotation and analysis. Due to overall low contribution of ambient RNA across
1284 all samples, CellRanger filtered counts were used throughout preprocessing and analysis.
1285
1286 Preprocessing for the scRNAseq data followed similar logic outlined in the Scanpy (v1.7.2) workflow
1287 tutorial[186]. Raw gene counts were normalised using *sc.pp.normalise_per_cell* function with
1288 counts_per_cell_after=1e4 and then a ln(x+1) transformation was applied using the *sc.pp.log1p* function.

1289 For any independent analysis shown, expression values reported are normalised then log-transformed as
1290 above.
1291
1292 ***Whole embryo: scRNAseq integration, batch correction and visualisation***
1293 For WE scRNA-seq integration, cell cycling genes were obtained from an public integrative fetal scRNAseq
1294 atlas[24], and the 542 intersecting cell cycle genes present in our dataset were excluded. Data were then log-
1295 normalised as described in '*scRNAseq gene expression visualisation using dotplots and heatmaps*'. Highly
1296 variable gene (HVG) selection was performed using the *sc.pp.highly_variable_genes* function from Scanpy
1297 (*min_mean=0.0125, max_mean=3, batch_key='haniffa_ID'*), and the top 10,000 highly variable genes
1298 HVGs ranked by *'dispersions_norm'* were flagged. For dimensionality reduction and integration, we used
1299 the scVI module from scvi-tools (v0.19.0). The top 10,000 HVGs identified above were input into scVI as
1300 raw expression counts. The scVI workflow[182] was then followed, with biological replicate and 10x chemistry
1301 taken as categorical covariates, and 256 nodes per hidden layer used. Following scVI integration, latent
1302 spaces were passed in the *sc.pp.neighbors* (*n_neighbors=30*) and *sc.tl.umap* functions from Scanpy. In order
1303 to evaluate optimum scVI model and number of neighbourhoods, a varying number of latent spaces (10, 20,
1304 30), batch sizes (1,024, 3,000, 5,000) and number of neighbours (15, 20, 25, 30) were scored for purity,
1305 homogeneity (*sklearn.metrics.homogeneity_score*), completeness
1306 (*sklearn.metrics.cluster.completeness_score*), normalised mutual information
1307 (*sklearn.metrics.cluster.normalized_mutual_info_score*) and global silhouette distance
1308 (*sklearn.metrics.silhouette_score*) for both biological replicate and chemistry. All metric outcomes were
1309 screened but global silhouette distance for both biological replicate and chemistry informed the final model
1310 and neighbourhood decision (20 latent spaces, 30 neighbours, batch size of 3,000). Finally, we utilised the
1311 uniform manifold approximation (UMAP) algorithm from the *sc.tl.umap* function from Scanpy for
1312 visualisation.
1313
1314 For annotation purposes, dimensionality reduction and integration as described above was independently re-
1315 run for cells and genes from each of 12 embryo grid-like anatomical tissue sections, and with the following
1316 differences: cell cycle genes were not removed, only top 6,000 HVG by dispersion were retained for scVI
1317 integration, default scVI parameters were utilised. For section-wise dimensional reduction, the optimised

1318 number of neighbours (between 15, 20, 25, and 30) were assessed using *sklearn.metrics* for silhouette
1319 scoring, with the best-performing number of neighbours (NNs) selected as follows: 15NNs for sections 2, 7,
1320 11; 20NNs for sections 3, 6, 9; 25NNs for sections 1, 4, 10; 30NNs for sections 5, 8, 12.
1321
1322 ***Whole embryo: scRNAseq manual annotation - clustering and differential expression testing***
1323 For manual annotation, clustering was performed independently on each of the 12 section-wise embryo
1324 scRNAseq objects using the Scanpy *sc.tl.leiden* function with a resolution parameter of *res=3.* Subclusters
1325 were then generated within each major cluster with a resolution parameter of *res=1*. Manual annotation was
1326 performed by a combination of literature markers, differentially expressed genes (DEGs) between major
1327 clusters using the *sc.tl.rank_genes_groups* function from Scanpy for a two-sided Wilcoxon rank-sum test
1328 (*corr_method='benjamini-hochberg'*) and information from the community data CellTypist model. DEG
1329 output was filtered for significant adjusted p-values (<0.05) and the top 100 significant DEGs per category
1330 were then ranked by log fold change. The resultant manual annotations were used for evaluation of QC
1331 metrics as described above (i.e., low quality cells according to mitochondrial reads, doublets, and maternal
1332 contaminants were removed), then these annotations were refined through the construction of a reference
1333 human embryonic dataset (described next).
1334
1335 ***HDCA: Integration and annotation harmonisation of the final integrated community and whole embryo***
1336 ***data***
1337 The HDCA dataset is the combination of the Whole Embryo and the Community data consisting of 4,679,782
1338 cells, split into 12 level 2 systems based on the initial annotated cell types (**Fig. 1c**). For each of the systems,
1339 initial round of annotation harmonisation of public data and embryo data were performed as follows: 4,000
1340 most highly variable genes (HVGs) were identified using *scanpy.pp.highly_variable_genes* with a
1341 concatenated variable of study, sample state (fresh or frozen), single cell or single-nuclei, and sequence
1342 technology as the batch key (*batch_key*). Batch correction was performed using scVI with default parameters
1343 using the above-mentioned batch key. kNN graphs were constructed using *scanpy.pp.neighbors* with k=15
1344 and the scVI latent space. The underlying kNN was used to cluster the cells using Leiden community
1345 detection using *scanpy.tl.leiden* (*resolution* = 1.0).
1346

1347 For the aerodigestive, audiovisual, endocrine, cardiovascular, hematopoietic, integumentary,
1348 undifferentiated germ cells, yolk sac and urogenital systems, we adopted a cluster centric embryo-
1349 community label harmonisation approach. The 6PCW embryo cell types at level 3 (broad) and lvl4 (fine)
1350 were predicted using a distance weighted kNN classifier (*sklearn.neighbors.KNeighborsClassifier*) with
1351 k=10 nearest neighbours from the community data. Clusters that were largely dominated by cells from
1352 community data, the whole embryo cells of those clusters were assigned the kNN-classifier predicted cell
1353 types and clusters where cells were predominantly embryo data, the community data cells were assigned the
1354 whole embryo cell type. For clusters where both community and embryo data were proportionally well
1355 represented, we adopted a manual approach with expert guidance for cell type harmonisation.
1356
1357 Cells of the three distributed systems namely the mesenchyme (**Fig 3**), vascular and lymphatic endothelium
1358 (**Fig 4**) and peripheral nervous system (**Fig 5**) were of particular focus in our manuscript. Cells were assigned
1359 to these systems based on the initial round of refined cell type annotations derived from both community
1360 datasets and de novo data. Harmonisation of the annotations across the community data and the embryo data
1361 were then performed with expert guidance. For each of these three groupings the HVGs and cluster
1362 resolutions were not treated as a fixed or standard parameter as for other systems above, but instead were
1363 selected in a data-driven manner based on whether clusters represented biologically meaningful and
1364 sufficiently pure cell populations (res; vasculature = 4.0, peripheral nervous system = 1.0, mesenchyme =
1365 3.0). For HVG used per system, please see **Supplementary Table 15**. Cell type identities were assigned
1366 using canonical marker genes (**Supplementary Table 9, 16**) and also informed by integrating the suspension
1367 and Xenium spatial data described below in section *Whole embryo: Xenium preprocessing and analysis*.
1368 After initially clustering and annotating each system independently, the refined cell types were collated and
1369 iteratively refined to assess and merge equivalent populations across systems. System-specific clustering
1370 also enabled us to more accurately resolve cells that had previously been assigned to broader categories. For
1371 example, within the mesenchymal system, higher-resolution clustering and annotation identified a subset of
1372 cells exhibiting vascular endothelial transcriptional profiles, which were subsequently reassigned to the
1373 vascular endothelial system. Clusters with less than 10 cells after refinement were assigned to k nearest
1374 neighbour's celltype. For celltypes from the other systems, the initial unbiased clustering framework
1375 provided a common reference structure for harmonising the community datasets with the de novo Whole

1376 Embryo data and for evaluating concordance between existing annotations across studies. During cross-
1377 system refinement, previously published annotations were retained where they were unambiguous and
1378 provided higher-resolution labels than those obtained from unbiased clustering. For example, in the
1379 hematopoietic system, a myeloid lineage cluster in which the majority of cells were annotated as
1380 Macrophage_LYVE1_High by Suo Dann et al., but which also contained cells with more generic annotations
1381 such as Macrophage from Zhang et al., was annotated as MACROPHAGE_LYVE1_HIGH. Central nervous
1382 system cells retained the original fine-grained annotations from Braun et al., as these annotations were
1383 informed by spatial context, a key variable that would not otherwise be captured in the suspension
1384 transcriptomic latent space. The final annotation spans across 12 systems, 69 broad cell types and 454 refined
1385 cell types and are provided in **Supplementary Table 9.** Post harmonisation and cross system refinement,
1386 concordance between the HDCA refined cell type annotations and the original author-provided annotations
1387 was assessed using Normalised Mutual Information (NMI; *sklearn.metrics.normalized_mutual_info_score*).
1388 This yielded an NMI score of 0.85, indicating strong agreement between the two annotation schemes. The
1389 top 50 differentially expressed genes ranked by z-scores and padj < 0.05, for each of the refined cell types is
1390 provided in **Supplementary Table 16**. The refined cell types that were expert annotated in the three
1391 distributed systems (the mesenchyme, vascular and lymphatic endothelium and peripheral nervous system)
1392 are listed in **Supplementary Table 17.**
1393
1394 To integrate the entire public and embryo data after per-system annotation harmonisation as above, we used
1395 the above-mentioned *batch_key* to identify 12,288 HVGs. Highly variable genes were selected in a batch-
1396 aware manner using *scanpy.pp.highly_variable_genes*, with dispersion normalized within each batch and
1397 aggregated globally to mitigate batch-driven effects. The top 12,288 genes ranked by global normalized
1398 dispersion were retained for downstream analysis (**Supplementary Table 15**). We trained scVI using a two-
1399 layer network (2,048 hidden units per layer), a 28-dimensional latent space, a dropout rate of 0.1, and a zero-
1400 inflated negative binomial (ZINB) gene likelihood with gene-wise dispersion, observed library size, and a
1401 normal latent prior. Batch effects were accounted for by including the *batch_key* as a covariate in scVI
1402 training, to generate the latent space. The above parameters were selected following a limited hyperparameter
1403 search constrained by the size of the dataset. We evaluated combinations of hidden layer sizes (512, 1024,
1404 and 2048 units), latent dimensionality (26, 28, and 30), number of hidden layers (2 and 3), and likelihood

models (negative binomial and zero-inflated negative binomial), using a dropout rate of 0.1. Model performance was assessed using scIB metrics computed on the initially predicted cell type annotations. The latent space was used to generate the kNN graph and this graph was used to generate the UMAP.

scRNAseq gene expression and UMAP visualisation

To visualise gene expression across dotplots, heatmaps or UMAPs, raw gene expression count matrices were subset to cell types of interest, and were then normalised using the Scanpy *sc.pp.normalise_per_cell* function with counts_per_cell_after=1e4. A ln(x+1) transformation was then applied to the normalised counts using the *sc.pp.log1p* function. Gene expression was then visualised with either the *sc.pl.dotplot*, *sc.pl.heatmap* or *sc.pl.umap/sc.pl.embeddings* functions. To increase interpretability of plots, the *vmax* argument was sometimes used to apply an upper limit to the color scale or scaled using *standard_scale='var'* (where used, noted within GitHub scripts).

scRNAseq cluster entropy calculation

To assess donor diversity across the atlas (**Fig. S2b**), within each high-resolution leiden cluster (resolution = 3), we calculated a donor entropy score. Clusters with a greater mix of donors have higher entropy scores, while clusters dominated by a single donor have lower scores. Entropy was computed using the Shannon entropy formula (base-2 logarithm) based on the proportion of cells contributed by each donor within a cluster. Entropy scores were normalized to range from 0 to 1 for comparability across clusters with different numbers of donors. The normalized entropy scores were then mapped back to individual cells for visualization.

scRNAseq cell-cell communication analysis

For the cross-system analysis within the Dorsal Root Ganglion (DRG), cell–cell communication analysis was performed using LIANA+[187,188] v1.5.0 on the spatially resolved Xenium dataset restricted to DRG cells found in **Fig. S24d**. Ligand–receptor interactions were inferred using the *rank_aggregate* function, which aggregates evidence across the consensus resource from cells grouped by cell types. Only ligands and receptors expressed in at least 10% of cells within each group were considered (*expr_prop = 0.1*), with a minimum of 5 cells per cell type and 100 permutations used to assess interaction specificity; raw expression

1434 values were retained. To incorporate spatial information, pairwise mean Euclidean distances between all
1435 cell-type pairs were calculated from a precomputed cell–cell distance matrix. Interactions were further
1436 filtered to retain only those with statistically significant CellPhoneDB-derived p-values (≤ 0.05).

1437 ***scRNAseq gene set enrichment analysis***

1438 Gene set enrichment scores for the PNS system were computed using the Univariate Linear Model (ULM)
1439 method implemented in the decoupleR[189] Python package (v2.1.1) in **Fig. 5m**, and **Fig. S23f, S23j, S24e**.
1440 For the enrichment of neural crest cells in skin organoids, curated marker gene signatures were used for NCC
1441 (*SOX10*, *FOXD3*), glial (*MPZ*, *PLP1*, *EDNRB*, *ERBB3*, *CDH19*), and neuronal (*ELAVL4*, *ELAVL3*, *PRPH*,
1442 *TUBB3*) transcriptional programs. To further resolve Schwann cell states, glial enrichment analyses
1443 additionally incorporated subtype-specific signatures for non-myelinating Schwann cells (*SCN7A*, *GFRA3*,
1444 *NCAM1*, *NRN1*) and myelinating Schwann cells (*MBP*, *PRX*, *PMP22*, *MAG*). To assess downstream receptor
1445 signalling activity, pathway gene sets were compiled from KEGG MEDICUS, Reactome, Gene Ontology
1446 Biological Process (GOBP), and BioCarta databases, including pathways associated with 'WNT5A–ROR
1447 signalling', 'NCAM-mediated neurite outgrowth', 'L1CAM interactions', 'ERBB2 signalling', 'CXCR4
1448 signalling', and 'HGF-MET-PI3K signalling'. In addition, ligand enrichment per niche was assessed using
1449 curated gene sets derived from GOBP and WikiPathways annotations, including 'Neuron migration',
1450 'Neuron maturation, Neural crest cell migration during development', 'Sympathetic ganglion development'
1451 and 'Peripheral nervous system neuron differentiation'. Enrichment scores were calculated at single-cell
1452 resolution to enable comparative analysis of lineage-specific programs and niche-derived ligand activity.

1453 Arterial and venous module scores (**Fig. 4b**) were computed using the *sc.tl.score_genes* function in Scanpy.
1454 Arterial scores were calculated based on the gene set *CXCL12, GJA5, EFNB2, HEY1, DLL4, SOX17,*
1455 *NOTCH4*, and *NRP1*, whereas venous scores were derived from *TLL1, ADAMTS18, SELP, NR2F2, EPHB4,*
1456 *APLNR, VCAM1,* and *NTRK2*. Adult gene program (**Fig. 4c**) scores were computed using the same approach,
1457 using genes derived from lists of DEGs from Supplementary Table 6 of a previously published scRNA-seq
1458 atlas of human adult endothelial cells[94]. To ensure robustness, we performed 20 randomised subsampling
1459 iterations. In each iteration, cells were randomly sampled from vascular and lymphatic cell types excluding
1460 the query population, matching the total number of cells in the query group. Scores were recalculated for

1461 each subsample to assess stability and minimise sampling bias. Linear Generalised Additive Model (GAM)
1462 with four splines was used to fit the scores.
1463

1464 ***Whole embryo: multiome sample acquisition and processing***

1465 Single nuclei were isolated from the remaining frozen cells from F137, F147 and F158 embryos that were
1466 not used for scRNAseq (https://www.protocols.io/view/nuclei-isolation-from-frozen-tissue-
1467 8epv5x6wjg1b/v1). Briefly, frozen cells were homogenised in homogenisation buffer (250 mM sucrose,
1468 25 mM KCl, 5 mM MgCl2, 10 mM Tris-HCl, 1 μM dithiothreitol, 1× protease inhibitor, 0.4 U μl−1
1469 RNaseIn, 0.2 U μl−1 SUPERaseIn and 0.1% Triton X-100 in nuclease-free water) using a pre-chilled
1470 Dounce homogeniser. Cells were left for 5 min on ice and then they were initially homogenised with
1471 a loose-fitting pestle "A" (4-8 strokes) followed by a tight-fitting pestle "B" (2-4 strokes). Homogenate
1472 was filtered through a 40μm cell strainer (CellTrics, Sysmex) and centrifuged for 6 min at 500 g
1473 (acceleration 0, deceleration 0) to obtain nuclei pellets. Pellets were resuspended in 500 ml of wash
1474 buffer (1× PBS, 5% BSA and 40 U μl−1 Protector RNaseIn) for 2 minutes on ice. After visualisation and
1475 assessment for viability under the microscope, nuclei suspensions were centrifuged again for 3 min at 500 g
1476 and the supernatant was carefully removed. Nuclei were manually counted and concentrations were
1477 estimated using an Improved Neubauer hemocytometer. All suspensions were maintained on ice until the
1478 transposition step.
1479

1480 ***Whole embryo: multiome library preparation, sequencing and alignment***

1481 Nucleus suspensions were loaded at a concentration targeting 10,000 nuclei per channel. Nuclei were
1482 subjected to transposition before GEM generation and barcoding. All embryos were processed using the
1483 Chromium Next GEM Single Cell Multiome ATAC + Gene Expression kit (10x Genomics) according to the
1484 manufacturer's instructions. Gene expression and ATAC libraries were prepared using proprietary 10x
1485 Genomics reagents and sequenced on the Illumina NovaSeq6000 S4, targeting a minimum depth of 50,000
1486 UMIs per cell. Multiome data (sn-RNA-seq and sn-ATAC-seq) were aligned to GRCh38-2020-A reference
1487 using cellranger-arc v2.0.2 pipeline with the default parameters from 10X CellRanger suite.

1488 ***Whole embryo: multiome snRNAseq QC preprocessing, annotation and visualisation***

1489 Raw counts from whole embryo multiome single nuclear RNA-seq (sn-RNA-seq) samples (**Supplementary**
1490 **Table 5**) with 467,575 cells and 36,601 genes were processed using Scanpy v1.11.2[186] (anndata v0.12.0)
1491 framework. Quality control metrics were calculated per cell, including total counts, and percentage of
1492 mitochondrial, ribosomal and hemoglobin counts. Cells were filtered to remove low-quality and doublets
1493 using the following thresholds:cells with >12,500, <200 detected genes and mitochondrial fractions > 5%
1494 were excluded. Genes expressed in fewer than 3 cells were removed. Counts were library-size normalised
1495 (shifted algorithm) using pp.normalize_total (target_sum=None), followed by log-transformation. Highly
1496 Variable Genes (HVGs) were identified using the Seurat v3 flavor with embryo_ID specified as the batch
1497 key, and the top 2,000 HGVs were retained. Dimensionality reduction was performed by PCA, followed by
1498 computing neighbours, UMAP and Leiden clustering (resolution parameter=1). Cells annotated as glia,
1499 neurons, neuronal progenitors were extracted based on established marker gene expression. The dataset was
1500 subset to these cells encompassing central nervous system (CNS) and Peripheral nervous system cells (PNS)
1501 populations for further subclustering. 3 Clusters (10, 12 and 13) were identified as PNS populations using
1502 known PNS markers and some CNS markers as negative markers (see Github) and were extracted for
1503 detailed annotation (**Fig. S8**). The PNS cells were subclustered in a new embedding space and leiden
1504 clustering of resolution = 0.5 identified 9 clusters corresponding to: Autonomic NCCs (Neuronal Crest cells)/
1505 Schwann cell precursors (SCPs), Sensory NCCs/SCPs, sensory neurons, sensory neuronal progenitors,
1506 sensory neurons (epibranchial placode), sensory neurons (otic placode), Autonomic neuronal progenitors,
1507 NCC/SCP-derived progenitors, NCC/SCP-derived mesenchyme, Melanocyte precursors. Cluster identities
1508 were assigned using known lineage markers and differential expression analysis (**Fig. S8c**; **Supplementary**
1509 **Table 9**). Final cell type labels were retained for cross-modality integration with snATAC-seq data, enabled
1510 by shared barcodes from the 10x Genomics multiome technology.

1511

1512 ***Whole embryo: multiome ATAC-seq preprocessing, peak calling and visualisation***

1513 snATAC-seq data were processed using cellranger-arc analytical suite and further filtered, and analysed
1514 downstream using SnapATAC2[190] v2.8.0 framework. Cells were retained if they met the following criteria,
1515 >1,000 unique fragments and TSS enrichment score >=7 (stricter than the default threshold of 5). In order
1516 to investigate the dataset further and assess the batch effect across embryos in the ATAC modality we

1517 processed to feature selections and dimensionality reduction. Feature selection was performed using 50,000
1518 genomics features prior to performing dimensionality reduction. For dimensionality reduction we used
1519 Laplacian Eigenmaps via spectral embedding implemented in the SnapATAC2 framework (snap.tl.spectral),
1520 followed by UMAP computation and visualisation with default parameters. After preprocessing and filtering,
1521 the whole embryo snATAC-seq dataset comprised 521,423 cells and 606,219 features. Since 10X multiome
1522 technology provides matched RNA and ATAC profiles from the same nuclei, snRNA-seq-derived
1523 annotations were directly transferred in ATAC modality using shared cell barcodes. From the full snATAC-
1524 seq dataset, 22,063 cells overlapped with PNS-annotated cell types defined in the snRNA-seq and were
1525 selected for downstream chromatin accessibility analysis. To ensure comparable peak detection across PNS
1526 cell populations, prior to peak calling cells were downsampled to a maximum of 2,000 cells per cell type.
1527 Cell types with less than 2,000 cells were not downsampled. This approach ensures more comparable and
1528 accurate peak identification across cell populations. Peak calling was performed using the MACS3[191] method
1529 implemented in the snapATAC2. After peak calling, coverage tracks were generated and exported in bigWig
1530 format using snap.ex.export_coverage function. The genomic tracks were then visualised in the Interactive
1531 Genomics Viewer (IGV) to examine chromatin accessibility patterns at specific marker gene loci across PNS
1532 populations (**Fig. S8**).
1533
1534 ***Whole embryo: multiome cell2fate analysis***
1535 We quantified spliced and unspliced counts with Velocyto (version v0.17.17) using GRCh38 v3. RNA
1536 velocity modules were separately modelled for PNS sensory and autonomic lineages using level 3 cell type
1537 annotations with cell2fate[153] (v.1.0.0, https://github.com/BayraktarLab/cell2fate) for **Fig. S23e**. Cell types
1538 outside lineages or with low cell numbers (parasympathetic neuronal, enteric glia, chromaffin cells,
1539 melanocyte precursors) were not considered for analysis. We used 3,000 highly-variable protein-coding
1540 genes (19,273 genes, GRCh38, HUGO Gene Nomenclature Committee, 2025-10-02 version) with at least
1541 10 detected counts, implementing the default dynamical model with embryo as batch key ("haniffa_ID") and
1542 8 modules. Model training was performed on an NVIDIA A100-SXM4-80GB GPU for 200 epochs until
1543 convergence with a default learning rate of 0,01. Module time estimates and activation/repression states were
1544 sampled from the posterior distribution (default 30 samples), and module-specific top genes, TFs, and GO-
1545 terms were calculated using *get_module_top_features* function.

1546 ***Whole embryo: Visium sample processing***

1547 A single embryo at Carnegie stage 17 (6 post conception weeks; PCW), sample SOB26, was formalin fixed
1548 and embedded in a paraffin (FFPE) block (**Supplementary Tables 1, 5**). One section at 4µm thickness was
1549 taken along the sagittal plane through the centre of the embryo (section_121) and through the left upper and
1550 lower limbs (section_221) using a microtome and collected on Superfrost plus microscope slides. 2 slides
1551 (section_121, section_221) were then dried for 3 hours at 42°C then placed in desiccant to avoid moisture.
1552 First, deparaffinization and Haematoxylin and Eosin staining were performed according to the standard 10X
1553 CytAssist workflow (CG000520). Slides were then imaged at 40X magnification on a Hamamatsu
1554 Nanozoomer. These slides were then taken forward for Visium analysis by destaining and decrosslinking
1555 the stained sections according to the standard 10X CytAssist workflow (CG000520) then proceeding
1556 immediately to perform the Visium CytAssist Spatial Gene Expression according to the standard 10X
1557 CytAssist workflow (CG000495). In brief this includes, probe hybridisation and probe ligation followed by
1558 CytAssist enabled probe release and extension before pre-amplification and cleanup. For these samples, the
1559 Visium_Human_Transcriptome_Probe_Set_v2.0_GRCh38-2020-A was used. Next, the samples were taken
1560 forward for probe-based library construction and quality control was performed using the Agilent
1561 Bioanalyzer High Sensitivity chip. 4 libraries were pooled at a time and sequenced down one lane of an
1562 Illumina Novaseq SP flow cell with the following run parameters: read 1: 28 cycles; i7 index: 10 cycles; i5
1563 index: 10 cycles; read 2S: 50 cycles.

1564

1565 ***Whole embryo: manual annotation of H&E images, and visualisation on Visium spatial data***

1566 Manual annotations of H&E section_121 from embryo SOB26 were performed by the Human
1567 Developmental Biology Resource (HDBR) team using the OMERO.iviewer (v0.16.0) platform. The original
1568 high-resolution images were hosted on the OMERO Plus server (https://omeroplus.sanger.ac.uk/) provided
1569 by Glencoe Software and maintained by the CellGenIT group at the Wellcome Sanger Institute. Regions of
1570 interest (ROIs) were outlined using the Draw Polygon tool within OMERO.iviewer to assign labels to known
1571 anatomical features.

1572 The resulting annotations were subsequently integrated with Visium spatial transcriptomics data using the
1573 ann2SR.py script from the OMERO_tools repository (https://github.com/cellgeni/OMERO_tools). This

1574 script retrieves the polygon coordinates of annotated regions from the OMERO server and combines them
1575 with the spatial coordinates of Visium spots. Based on geometric overlap, each Visium spot is assigned to
1576 annotated anatomical regions in both a categorical manner (discrete region label) and a continuous manner
1577 (proportional overlap with annotated regions) (**Fig. S7b**). The script outputs a CSV file containing spot (cell)
1578 identifiers together with the corresponding annotation assignments.

1579 ***Whole embryo: Visium gene expression analysis***

1580 Spatial transcriptomics Visium data (section_121) was mapped using SpaceRanger (v2.1.0) using GRCh38-
1581 2020-A reference. Spots were filtered to retain spots with total RNA counts >2000 and <200,000. Low count
1582 genes were then removed using the *cell2location.utils.filtering.filter_genes* function with the following
1583 parameters: cell_count_cutoff=5, cell_percentage_cutoff2=0.15, nonz_mean_cutoff=1.11. Genes were
1584 filtered to the set shared with scRNAseq data. This yielded Visium data comprising 2 sections (section_121
1585 and section_221) with 9,043 spots, and 17,824 genes. Only section_121 was used for downstream analysis
1586 comprising 7,778 spots. For plotting gene expression within the DRG region per Visium voxel (**Fig. 5j**) the
1587 spots corresponding to the DRG region were manually drawn as a region of interest (ROI) using the 10x
1588 Genomics Loupe Browser and the spatial coordinates of selected DRG spots were exported as CSV files.
1589 From the selected spots, the x and y centroid coordinates of each spot in pixel units (high resolution) were
1590 converted to micrometres (µm) using the Visium platform scale factor to represent true physical distance in
1591 tissue space.For the distance calculation from the DRG spots, the spatial proximity between spots was
1592 calculated by constructing a *k*-dimensional tree (k-d tree) using the full resolution spatial coordinates of DRG
1593 spots and *scipy.spatial.cKDTree(k=1) function*. The nearest Euclidean distance between each spot in the
1594 section_121 to its nearest DRG_spots was computed by querying each against the constructed KDTree and
1595 resulting distance metric for each spot (in µm) was stored as a per-spot metadata variable. To visualise the
1596 results, mean expression of each gene set was aggregated per cell and binned by distance from the DRG
1597 across a fixed number of equally spaced bins (n=10); per-bin means and standard errors were computed for
1598 each sample, then smoothed using a centered rolling window (window size=5) to produce continuous
1599 expression-by-distance profiles with 95% confidence intervals.
1600
1601

1602 ***Whole embryo: Xenium sample acquisition and processing***

1603 FFPE tissue samples from sample SOB25 (Carnegie stage 16) and SOB26 (Carnegie stage 17) were
1604 sectioned onto Xenium slides (10x Genomics 3000941) at a thickness of 4µm and stored at RT in desiccant
1605 (**Supplementary Table 1, 5**). These were then processed with the Xenium Prime workflow as per the
1606 manufacturer's protocols: CG000580 for pretreatment including deparaffinization and decrosslinking; and
1607 CG000760 for probe hybridisation, rolling circle product generation, segmentation marker staining,
1608 autofluorescence quenching, and nuclear counterstaining. Slides were loaded onto a Xenium Analyzer for
1609 automated cyclic fluorescent labelling and imaging as per the manufacturer's protocol CG000584, with
1610 Xenium Onboard Analysis software v3.0.0.15 or v3.1.0.4. Data were exported in the vendor format of
1611 ome.tif images (segmentation stain images including DAPI-stained nuclei), .h5 cell-gene count matrices,
1612 and .csv cell and transcript data.

1613

1614 ***Whole embryo: Xenium preprocessing and analysis***

1615 We used default 10x cell segmentation, utilising segmentation marker staining. 10x Genomics Xenium data
1616 was filtered to exclude cells with <10 genes per cell. We then used SCANVI (number of layers: 1; number
1617 of latent dimensions: 10) using labelled scRNA-seq and unlabelled Xenium data, restricted to ~5000 genes
1618 found in both datasets. Batch-aware HVG selection was then performed by setting the batch key to sample
1619 ID. We selected 2,000 HVGs as features for the integration. The batch key was sample id and a categorical
1620 covariate of technology (scRNA vs spatial) was used. We used this initial integration to broadly label groups
1621 in Xenium data (e.g. vascular, peripheral nervous system). We then repeated feature (HVG) selection and
1622 integration by lineage. For each lineage, we calculated the kNN graph and performed Leiden clustering. We
1623 labelled clusters based on a combination of scRNA-seq annotations and spatial location, with spatial
1624 positions helping to refine names for populations. We defined *FRZB*$^{+}$ progenitors as populations that had an
1625 uncertain identity after expert annotation. In contrast, other progenitors, such as dermal fibroblast progenitors
1626 and limb fibroblast progenitors, had been assigned a clear identity after expert annotation. For the
1627 predominantly mesenchymal cluster, we repeated the by-lineage integration separately for each group due
1628 to the increased complexity of this group (>3 million cells overall). These joint integrations, and the same
1629 marker genes used for scRNA-seq and Xenium-5k data, are shown in **Appendix B**.

1630

1631 ***Whole embryo: Niche identification analysis***

1632 We ran NicheCompass[51] (v0.3.0), a graph neural network-based approach, to obtain niche embeddings. We
1633 split whole embryo sections into 8 reference and 3 query sections (**Supplementary Table 5)**, to illustrate
1634 how our reference can be used to obtain niche embeddings for future datasets. Spatial neighbourhood graphs
1635 were constructed using Squidpy (8 nearest neighbours), and curated gene programs were assembled using
1636 NicheCompass. NicheCompass was trained on the reference dataset with default parameters and batch key
1637 as sample_id for 100 epochs. Niche annotations were based on histopathological annotations of whole
1638 embryos and differentially expressed genes (**Supplementary Table 12**), calculated after log1p normalisation
1639 using Scanpy's rank_genes_groups function with a t-test. We also used niche cell compositions predicted
1640 from SCANVI. For query data, the trained model was subsequently fine-tuned on the query data and both
1641 datasets were integrated into a shared latent space. All weights were frozen except categorical covariates
1642 embedder weights (unfreeze_cat_covariates_embedder_weights == True). Fine-tuning was otherwise
1643 performed using the same training objective and hyperparameters as reference training. Low-dimensional
1644 embeddings were computed using UMAP after kNN construction (k=15). For querying disease genes across
1645 niches, we used a reported gene list for Mendelian genes for congenital cataracts[75].

1646

1647 ***HDCA mapping to skin organoid for mesenchyme analysis***

1648 We jointly integrated human skin organoid (unlabelled for integration) with our HDCA using scanVI. We
1649 included only populations labelled as fibroblast: organoid_Fibroblasts POSTN+, organoid_Pre-dermal
1650 condensate, organoid_Early fibroblasts FRZB+, organoid_Early fibroblasts HOXC5+, organoid_Dermal
1651 condensate, organoid_Dermal papilla, organoid_Adipocytes, organoid_Neuron progenitors,
1652 organoid_Neuron progenitors SPP1+, organoid_Neuroepithelial cells, organoid_Smooth muscle/Pericytes,
1653 organoid_Myoblasts, organoid_Myocytes, organoid_Early myocytes. We included samples for HVG
1654 selection (~2000 target genes in a batch-aware manner, using sample id as batch key) if they contain >1000
1655 cells. Genes associated with mitochondrial, ribosomal, hemoglobin, cell-cycle, hypoxia, and stress responses
1656 were removed. kNN graph construction (k=20) was performed followed by Leiden clustering. We identified
1657 Leiden clusters enriched for progenitor populations, subset the dataset to these clusters, and recomputed the
1658 k-nearest neighbor graph in the SCANVI latent space. UMAP embeddings were initialized using the PAGA

1659 graph to preserve global trajectory structure. SCANVI was used to transfer new annotations from reference
1660 atlases to organoid data, which we validated using the same marker genes used for the HDCA populations
1661
1662 ***Whole embryo: RNAscope in situ hybridization imaging***
1663 Human embryonic tissues for RNAScope were firstly fixed in 10% PFA overnight at room temperature, and
1664 then secondly fixed in Methacarn before they were processed using automated Excelsior AS (Thermo
1665 Scientific, A823-1005) to generate FFPE blocks. Sections were cut from the blocks at 8 µm using a manual
1666 rotary microtome (Leica microsystems RM2235, UK). Simultaneous RNA *in situ* hybridization was carried
1667 out using RNAscope Multiplex Fluorescent Reagent Kit v2 (Cat. No. 323100 and 323270, ACD) following
1668 the manufacturer's protocol for FFPE tissue mounted on slides, with pre-treatment steps of Hydrogen
1669 Peroxide and Protease reagents (Cat. No. 322381, ACD) and target retrieval reagents for 20mins at 95ºC
1670 (Cat. No. 322000, ACD). The following RNAscope probes were designed and purchased from ACD: Hs-
1671 PIEZO1 (Cat No. 485101-C3), Hs-PROX1 (Cat No. 530241) and VEGFR3 (Hs-FLT4) (Cat No. 552441-
1672 C2). The probes were fluorescently labelled using OPAL 520 (Cat. No. FP1487001KT, PerkinElmer), OPAL
1673 570 (Cat. No. FP1488001KT, PerkinElmer) and OPAL 650 dyes (Cat. No. FP1496001KT, PerkinElmer) for
1674 direct visualization under automated laser-scanning microscopy (Zeiss CellDiscoverer7, Carl Zeiss,
1675 Germany) (**Fig. 4l**).
1676
1677 ***Whole embryo: RareCyte immunofluorescence staining and imaging***
1678 The following steps were performed at room temperature unless otherwise noted. Formalin fixed tissue from
1679 embryo sample SOB26 (**Supplementary Table 1**) embedded in paraffin (FFPE) were sectioned at 4µm
1680 thickness with a microtome (Leica RM2235) and mounted onto superfrost slides (Fisher Scientific
1681 12312148) before drying at 60°C for 60 minutes to ensure adhesion. The mounted sections were then
1682 deparaffinized and antigen retrieval was performed using the BioGenex EZ-Retriever system for 95°C for 5
1683 minutes then 107°C for 5 minutes. Slides were then bleached with AF Quench Buffer (4.5% H2O2 / 24 mM
1684 NaOH in PBS) to remove autofluorescence. Quenching was performed using a UV transilluminator using
1685 the high setting for 60 minutes with a strong white light exposure followed by 30 minutes using the 365 nm
1686 high setting. Slides were then rinsed with 1X PBS before incubating in 300 µl of Image-iT™ FX Signal
1687 Enhancer (Thermo Fisher I36933) for 15 min and rinsed again. The primary labelled antibodies Ki67

1688 (RareCyte, 52-1013-501, fluorophore ArgoFluor 555L, clone D3B5) and Pan-CK (RareCyte, 52-1015-801,
1689 fluorophore ArgoFluor 845, clone C11/AE1/AE3) were prediluted according to manufacturer's
1690 recommendations and added to the tissue then incubated for 120 minutes in the dark on a humidity tray. The
1691 antibodies used are provided in **Supplementary Table 3**. Slides were then washed with a surfactant wash
1692 buffer then 300 µl of nuclear staining in goat diluent was added to the slide, before incubating in the dark
1693 for 30 minutes on a humidity tray. Slides were washed and placed in 1X PBS before adding a coverslip with
1694 ArgoFluor mount media and left in the dark at room temperature overnight to dry. For imaging, the RareCyte
1695 Orion microscope was used with a 20X objective lens (**Fig. S1f**). Scans were performed using Imager and
1696 relevant acquisition settings were applied using the software Artemis. Slides were subsequently transferred
1697 to -20°C for extended storage.
1698
1699 ***Whole embryo: MACSIMA immunofluorescence staining and imaging***
1700 For MACSIMA immunofluorescence imaging of the dorsal root ganglion (**Fig. 5i; Supplementary Table**
1701 **1**), a 7PCW human embryonic (sample SM74P7) was cryoprotected in a solution of 10% sucrose at 4°C for
1702 3 days. It was then embedded in a mixture of 7.5% gelatin, 10% sucrose, and 0.12 M phosphate buffer,
1703 followed by freezing in isopentane at -55°C and storage at -80°C until needed. The sample was cut into
1704 10µm sections using a cryostat (Leica CM3050) and placed on Superfrost slides for frozen sections.
1705 Immunofluorescence staining was performed using the MACSima™ Imaging Platform. This imaging
1706 system operates by iterative immunofluorescent staining, sample washing, imaging, and signal bleaching,
1707 using 3 fluorochrome-conjugated antibodies per cycle. The antibodies used are provided in **Supplementary**
1708 **Table 3**. Image datasets (stack of tiff images) were analyzed using the MACS iQ View Software Version
1709 1.3.2.
1710
1711 ***HDCA web portal integration with developmental disorders database***
1712 We downloaded the DDG2P[43] database from the G2P website (https://www.ebi.ac.uk/gene2phenotype,
1713 accessed on 16/10/2024). DDG2P gene-disease records are annotated with a confidence level which reflects
1714 the level of evidence supporting the association. Records with 'limited' confidence were removed, keeping
1715 only associations with 'definitive', 'strong' and 'moderate' confidence. We then stored gene-disease pairs
1716 in a relational database and used the open-source software Strapi to expose a web API to allow our Cherita

web portal to query and retrieve the data dynamically by either gene or disease as per user exploration. Cherita, derived from the Malay word for "story", embodies the core purpose of the tool: enabling users to derive biological insights through their data. For more detailed information about its functionality and applications, please refer to the GitHub repository at *https://github.com/haniffalab/cherita-react*.

1721 **Acknowledgements**

1722 We are grateful to the donors and donor families for granting access to the tissue samples. We thank the
1723 Newcastle University Flow Cytometry Core Facility, Newcastle University Genomics Facility, Sanger
1724 Institute Cellular Genetics IT, CRUK CI Genomics Core Facility, and Newcastle upon Tyne NHS Trust
1725 NovoPath. We thank Mr Mohi Miah for wet-lab tissue processing assistance. We thank Dr Aidan Maartens
1726 and Dr Vicki Moignard for their writing contributions and suggestions. We also thank Martin Prete for his
1727 technical guidance on the use of the OMERO platform for anatomical annotations of H&Es.

1728

1729 **Funding**

1730 This research was funded in whole, or in part, by an MRC Human Cell Atlas award, Wellcome Human Cell
1731 Atlas Strategic Science Support (WT211276/Z/18/Z), the Wellcome Human Developmental Biology
1732 Initiative (WT215116/Z/18/Z), the Wellcome Trust (203151/Z/16/Z, 203151/A/16/Z, 226083/Z/22/Z),
1733 UKRI Medical Research Council (MC_PC_17230), HDBR (MRC/Wellcome MR/R006237/1) and Chan
1734 Zuckerberg Initiative Foundation (CZIF2022-007488). This work was supported by the de.NBI Cloud within
1735 the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum
1736 Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014,
1737 W-de.NBI-016, W-de.NBI-022). This publication is part of the Human Cell Atlas, thanks to members of the
1738 Developmental Biological Network (https://www.humancellatlas.org/biological-networks/development-
1739 biological-network/). M.H. is funded by Wellcome (WT107931/Z/15/Z, WT223092/Z/21/Z, WT206194,
1740 and WT220540/Z/20/A), the Lister Institute for Preventive Medicine, and NIHR and Newcastle Biomedical
1741 Research Centre. F.P. is funded by Wellcome (226579/Z/22/Z), Kennedy Trust for Rheumatology Research
1742 (KENN 21 22 11), Versus Arthritis (22981) and the Lister Institute for Preventive Medicine. P.H. is
1743 supported by start-up funding from the Department of Pathology, an endowment from Bakar ImmunoX and
1744 the Esther Simon Memorial Fund from the Research Evaluation and Allocation Committee at UCSF. H.V.F
1745 is funded by Wellcome (226083/Z/22/Z) and is also supported by the NIHR Cambridge Biomedical Research
1746 Centre (NIHR203312); C.F.W. is supported by the NIHR Exeter Biomedical Research Centre; the views
1747 expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and
1748 Social Care. DECIPHER is hosted by EMBL-EBI and funding for the DECIPHER project is provided by
1749 the Wellcome Trust (WT223718/Z/21/Z). I.S. is funded by Barts Charity (MGU045), and the British Skin

Foundation (004/RA/23). D.P. is supported by Fundação para a Ciência e a Tecnologia (2020.08715.BD) and CZI Next-Gen Pilot Project grant (2023-330280). S.W. is funded by the Royal Society (CDF/R1/241008), and supported by the NIHR and Newcastle Biomedical Research Centre. B.G. is supported by Wellcome (215116/Z/18/Z; 220379/B/20/Z; 226795/Z/22/Z; 309075/Z/24/Z). S.S. is supported by BBSRC grant BB/Y003179/1. D.J. is supported by a Wellcome Trust Accelerator Award (314710/Z/24/Z) and the NIHR Biomedical Research Centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London. S.MA. and P.O. were funded by the BHF (SP/13/5/30288 and RG/17/7/33217) and the UKRI Medical Research Council (MR/P011543/1). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing interests. S.A.T. is a scientific advisory board member of ForeSite Labs, OMass Therapeutics, Qiagen, Xaira Therapeutics, a co-founder and equity holder of TransitionBio and Ensocell Therapeutics, a non-executive director of 10x Genomics and a part-time employee of GlaxoSmithKline.

Author contributions

Conceptualisation: M.H. Project Administration: S.W, MS.V.B, V.R. Formal Analysis: MS.V.B, S.W, A.R, C.X, L.S, M.A.K, K.I, D.J, A.V.Po, MA.J, K.T, K.Ra, S.Mad, J.T.H.L, S.Mak, J.J, K.P.C, L.D, G.DA. Data Curation: MS.V.B, S.W, A.R, C.X, L.S, M.A.K, K.I, D.J, K.Ra, A.V.Po, E.K, M.A.C, J.J. Investigation: E.S, NJ.C, J.B, A.A, S.MA, R.A.B, K.Ro, B.O, C.H, J.E, E.P, N.H.G, D.M. Visualisation: E.S, A.R.F, D.B.L, C.A, E.K, M.A.C, H.V.F, C.F.W. Data Interpretation: S.W, A.R, C.X, L.S, M.A.K, E.S, K.I, D.J, A.R.F, NJ.C, A.V.Po, MA.J, K.T, K.Ra, J.K, A.L, J.E.G.L, V.K, S.L, J.T.H.L, A.A, E.U, T.P.S, V.L, F.T, R.A.B, B.O, A.M.C, H.J.W, N.H.G, P.H, M.W.M, B.T, L.F, D.P, C.M, R.E, S.Ba, V.R.K.S, K.K, J.C, S.S, I.S, M.D.L, S.MA, P.O, O.P, L.G, P.A, V.B.MS, F.J.T, F.P. Software: D.B.L, D.H, A.R, C.X, L.S, J.F, T.L, K.P, A.V.Pr, P.M. M.L. Resources: S.L, M.C, A.G, A.F, J.K, R.V.T, A.C, D.T, N.K.W, J.M. Funding Acquisition: M.H, O.A.B, S.A.T, K.B.M, B.G, E.R, S.Be. Supervision: M.H, O.A.B, F.P, F.J.T, S.A.T, K.B.M. Writing – Original Draft Preparation: S.W, C.X, L.S, D.J, F.P. Writing - Review & Editing: S.W, M.H. All authors reviewed and approved the manuscript.

Data and code availability

Data

Raw de novo whole embryo data will be available on EMBL-EBI ArrayExpress and BioImage Archive. All data accessions are currently private, and will be made available upon publication:

- scRNAseq raw FASTQ and counts (embryo F137), E-MTAB-11504
- scRNAseq raw FASTQ and counts (embryo F147), E-MTAB-11520
- scRNAseq raw FASTQ and counts (embryo F158), E-MTAB-11911

- Multiome raw FASTQ and counts (embryos F137, F147, F158), E-MTAB-16757

- Spatial transcriptomics Visium raw FASTQs, counts, images (embryo SOB26), E-MTAB-14785
- Spatial transcriptomics Xenium raw cell x gene counts post Xenium cell segmentation, images (embryos SOB25, SOB26), BioImage Archive, S-BIAD2878

The HDCA atlas will be accessible via an interactive web portal at https://cellatlas.io/studies/hdca/. This webportal is currently private and password protected, and will be made available upon publication. Through this portal, users will be able to visualise the dataset and download both raw and processed count matrices, as well as associated metadata and models.

Code

All code will be made publicly available at https://github.com/haniffalab/hdca-paper; the repository is currently private, and will be made available upon publication.

Additional repositories associated with this study include https://github.com/theislab/idtrack[181] for biological identifier mapping and harmonization across heterogeneous data sources, and

https://github.com/theislab/sctram[45], which provides the trajectory-fidelity benchmarking framework introduced in this study.

1807 **References**

1808 1. O'Rahilly, R. & Müller, F. *Developmental Stages in Human Embryos: Including a Revision of*
1809 *Streeter's 'Horizons' and a Survey of the Carnegie Collection*. (Carnegie Institute (WA), 1987).
1810 2. Lindsay, S. J. *et al.* HDBR Expression: A Unique Resource for Global and Individual Gene
1811 Expression Studies during Early Human Brain Development. *Front Neuroanat* **10**, 86 (2016).
1812 3. Belle, M. *et al.* Tridimensional Visualization and Analysis of Early Human Development. *Cell*
1813 **169**, 161–173.e12 (2017).
1814 4. Dawood, Y. *et al.* Human embryonic and fetal biobanking: Establishing the Dutch Fetal Biobank
1815 and a framework for standardization. *Dev Cell* **58**, 2826–2835 (2023).
1816 5. Haniffa, M. *et al.* A roadmap for the Human Developmental Cell Atlas. *Nature* **597**, 196–205
1817 (2021).
1818 6. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity
1819 and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* **32**, 1053–1058
1820 (2014).
1821 7. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309
1822 (2020).
1823 8. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem
1824 Cells. *Cell* **167**, 566–580.e19 (2016).
1825 9. Braun, E. *et al.* Comprehensive cell atlas of the first-trimester developing human brain. *Science*
1826 **382**, eadf1226 (2023).
1827 10. Holloway, E. M. *et al.* Mapping Development of the Human Intestinal Niche at Single-Cell
1828 Resolution. *Cell Stem Cell* **28**, 568–580.e4 (2021).
1829 11. Elmentaite, R. *et al.* Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional
1830 Links to Childhood Crohn's Disease. *Dev Cell* **55**, 771–783.e5 (2020).
1831 12. Cui, Y. *et al.* Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human
1832 Heart. *Cell Rep* **26**, 1934–1950.e5 (2019).
1833 13. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing

1834 Human Heart. *Cell* **179**, 1647–1660.e19 (2019).
1835 14. Kanemaru, K. *et al.* Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–
1836 810 (2023).
1837 15. Goh, I. *et al.* Yolk sac cell atlas reveals multiorgan functions during human early development.
1838 *Science* **381**, eadd7564 (2023).
1839 16. Popescu, D.-M. *et al.* Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019).
1840 17. Jardine, L. *et al.* Blood and immune development in human fetal bone marrow and Down
1841 syndrome. *Nature* **598**, 327–331 (2021).
1842 18. Stewart, B. J. *et al.* Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–
1843 1466 (2019).
1844 19. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of
1845 renal tumors. *Science* **361**, 594–599 (2018).
1846 20. He, P. *et al.* A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation
1847 and key regulators of epithelial fates. *Cell* **185**, 4841–4860.e25 (2022).
1848 21. Zhang, B. *et al.* A human embryonic limb cell atlas resolved in space and time. *Nature* **635**, 668–
1849 678 (2024).
1850 22. To, K. *et al.* A multi-omic atlas of human embryonic skeletal development. *Nature* **635**, 657–667
1851 (2024).
1852 23. Lawrence, J. E. G. *et al.* Single-cell transcriptomics identifies chondrocyte differentiation
1853 dynamics in vivo and in vitro. *Dev Cell* **60**, 3066–3084.e8 (2025).
1854 24. Suo, C. *et al.* Mapping the developing human immune system across organs. *Science* **376**,
1855 eabo0510 (2022).
1856 25. Reynolds, G. *et al.* Developmental cell programs are co-opted in inflammatory skin disease.
1857 *Science* **371**, eaba6500 (2021).
1858 26. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans.
1859 *Nature* **563**, 347–353 (2018).
1860 27. Park, J.-E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation.
1861 *Science* **367**, (2020).

1862 28. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional
1863 programs. *Nature* **572**, 67–73 (2019).
1864 29. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nat Med* **29**, 1563–
1865 1577 (2023).
1866 30. Domínguez Conde, C. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in
1867 humans. *Science* **376**, eabl5197 (2022).
1868 31. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell
1869 transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
1870 32. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding
1871 disease gene function. *Science (New York, N.Y.)* **376**, (2022).
1872 33. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes by integrating
1873 single-cell RNA-sequencing and human genetics. *Nat Genet* **54**, 1479–1492 (2022).
1874 34. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis.
1875 *Nature* **566**, 490–495 (2019).
1876 35. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**,
1877 496–502 (2019).
1878 36. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*
1879 **576**, 487–491 (2019).
1880 37. Niu, Y. *et al.* Dissecting primate early post-implantation development using long-term in vitro
1881 embryo culture. *Science* (2019) doi:10.1126/science.aaw5754.
1882 38. Zhai, J. *et al.* Primate gastrulation and early organogenesis at single-cell resolution. *Nature* **612**,
1883 732–738 (2022).
1884 39. Zeng, B. *et al.* The single-cell and spatial transcriptional landscape of human gastrulation and
1885 early brain development. *Cell Stem Cell* **30**, 851–866.e7 (2023).
1886 40. Xu, Y. *et al.* A single-cell transcriptome atlas profiles early organogenesis in human embryos.
1887 *Nat. Cell Biol.* **25**, 604–615 (2023).
1888 41. Cui, L. *et al.* Spatial transcriptomic characterization of a Carnegie stage 7 human embryo. *Nature*
1889 *Cell Biology* **27**, 360–369 (2025).

1890 42. Pan, J. *et al.* Spatiotemporal transcriptome atlas of human embryos after gastrulation. *bioRxiv*
1891 (2023) doi:10.1101/2023.04.22.537909.
1892 43. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P
1893 with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).
1894 44. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat*
1895 *Methods* **19**, 41–50 (2022).
1896 45. Inecik, K., Rose, A., Haniffa, M., Luecken, M. & Theis, F. J. Beyond Visual Inspection:
1897 Principled Benchmarking of Single-Cell Trajectory Representations with scTRAM. *bioRxiv*
1898 2025–2026 (2025).
1899 46. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat*
1900 *Biotechnol* **40**, 121–130 (2022).
1901 47. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
1902 48. Zhao, C. *et al.* A comprehensive human embryo reference tool using single-cell RNA-
1903 sequencing data. *Nature Methods* **22**, 193–206 (2024).
1904 49. Sthijns, M. M. J. P. E., LaPointe, V. L. S. & van Blitterswijk, C. A. Building Complex Life
1905 Through Self-Organization. *Tissue Eng Part A* **25**, 1341–1346 (2019).
1906 50. Sasai, Y. Cytosystems dynamics in self-organization of tissue architecture. *Nature* **493**, 318–326
1907 (2013).
1908 51. Birk, S. *et al.* Quantitative characterization of cell niches in spatially resolved omics data. *Nat.*
1909 *Genet.* **57**, 897–909 (2025).
1910 52. Parker, K. L. & Schimmer, B. P. Steroidogenic factor 1: a key determinant of endocrine
1911 development and function. *Endocr Rev* **18**, 361–377 (1997).
1912 53. Bouchard, M., Souabni, A., Mandler, M., Neubüser, A. & Busslinger, M. Nephric lineage
1913 specification by Pax2 and Pax8. *Genes Dev* **16**, 2958–2970 (2002).
1914 54. Steventon, B., Mayor, R. & Streit, A. Mutual repression between Gbx2 and Otx2 in sensory
1915 placodes reveals a general mechanism for ectodermal patterning. *Dev Biol* **367**, 55–65 (2012).
1916 55. Magaletta, M. E. *et al.* Integration of single-cell transcriptomes and chromatin landscapes reveals
1917 regulatory programs driving pharyngeal organ development. *Nat Commun* **13**, 457 (2022).

1918 56. Kameda, Y., Nishimaki, T., Chisaka, O., Iseki, S. & Sucov, H. M. Expression of the epithelial
1919 marker E-cadherin by thyroid C cells and their precursors during murine development. *J*
1920 *Histochem Cytochem* **55**, 1075–1088 (2007).
1921 57. Macchia, P. E. *et al.* PAX8 mutations associated with congenital hypothyroidism caused by
1922 thyroid dysgenesis. *Nat. Genet.* **19**, 83–86 (1998).
1923 58. Kamitani-Kawamoto, A. *et al.* MafB interacts with Gcm2 and regulates parathyroid hormone
1924 expression and parathyroid development. *J. Bone Miner. Res.* **26**, 2463–2472 (2011).
1925 59. Zou, D. *et al.* Patterning of the third pharyngeal pouch into thymus/parathyroid by Six and Eya1.
1926 *Dev Biol* **293**, 499–512 (2006).
1927 60. Lanigan, T. M., DeRaad, S. K. & Russo, A. F. Requirement of the MASH-1 transcription factor
1928 for neuroendocrine differentiation of thyroid C cells. *J Neurobiol* **34**, 126–134 (1998).
1929 61. Johansson, E. *et al.* Revising the embryonic origin of thyroid C cells in mice and humans.
1930 *Development* **142**, 3519–3528 (2015).
1931 62. Massalha, H. *et al.* A developmental cell atlas of the human thyroid gland. *Genomics* (2024).
1932 63. Westerlund, J. *et al.* Expression of Islet1 in thyroid development related to budding, migration,
1933 and fusion of primordia. *Dev Dyn* **237**, 3820–3829 (2008).
1934 64. Kusakabe, T., Hoshi, N. & Kimura, S. Origin of the ultimobranchial body cyst: T/ebp/Nkx2.1
1935 expression is required for development and fusion of the ultimobranchial body to the thyroid.
1936 *Dev Dyn* **235**, 1300–1309 (2006).
1937 65. Salzet, M. How the diffuse neuroendocrine system shapes health, homeostasis, and cancer. *Front*
1938 *Neuroendocrinol* **79**, 101216 (2025).
1939 66. Wang, W., Chan, E. K., Baron, S., Van de Water, T. & Lufkin, T. Hmx2 homeobox gene control
1940 of murine vestibular morphogenesis. *Development* **128**, 5017–5029 (2001).
1941 67. Ponnio, T., Burton, Q., Pereira, F. A., Wu, D. K. & Conneely, O. M. The nuclear receptor Nor-1
1942 is essential for proliferation of the semicircular canals of the mouse inner ear. *Mol Cell Biol* **22**,
1943 935–945 (2002).
1944 68. Salminen, M., Meyer, B. I., Bober, E. & Gruss, P. Netrin 1 is required for semicircular canal
1945 formation in the mouse inner ear. *Development* **127**, 13–22 (2000).

1946 69. Lu, Y. *et al.* Single-Cell Analysis of Human Retina Identifies Evolutionarily Conserved and
1947 Species-Specific Mechanisms Controlling Development. *Dev Cell* **53**, 473–491.e9 (2020).
1948 70. Zuo, Z. *et al.* Single cell dual-omic atlas of the human developing retina. *Nat Commun* **15**, 6792
1949 (2024).
1950 71. Ghiasvand, N. M. *et al.* Deletion of a remote enhancer near ATOH7 disrupts retinal
1951 neurogenesis, causing NCRNA disease. *Nat Neurosci* **14**, 578–586 (2011).
1952 72. Watanabe, S. *et al.* Prdm13 regulates subtype specification of retinal amacrine interneurons and
1953 modulates visual sensitivity. *J Neurosci* **35**, 8004–8020 (2015).
1954 73. Clark, B. S. *et al.* Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors
1955 as Regulating Mitotic Exit and Late-Born Cell Specification. *Neuron* **102**, 1111–1126.e5 (2019).
1956 74. Eldred, K. C. *et al.* Thyroid hormone signaling specifies cone subtypes in human retinal
1957 organoids. *Science* **362**, (2018).
1958 75. Shiels, A. & Hejtmancik, J. F. Inherited cataracts: Genetic mechanisms and pathways new and
1959 old. *Exp Eye Res* **209**, 108662 (2021).
1960 76. Lachke, S. A. *et al.* Mutations in the RNA granule component TDRD7 cause cataract and
1961 glaucoma. *Science* **331**, 1571–1576 (2011).
1962 77. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online
1963 Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic
1964 disorders. *Nucleic Acids Res* **43**, D789–98 (2015).
1965 78. Torregrossa, M. *et al.* Effects of embryonic origin, tissue cues and pathological signals on
1966 fibroblast diversity in humans. *Nat Cell Biol* **27**, 720–735 (2025).
1967 79. Buechler, M. B. *et al.* Cross-tissue organization of the fibroblast lineage. *Nature* **593**, 575–579
1968 (2021).
1969 80. Gopee, N. H. *et al.* A prenatal skin atlas reveals immune regulation of human skin
1970 morphogenesis. *Nature* **635**, 679–689 (2024).
1971 81. Qiu, C. *et al.* A single-cell time-lapse of mouse prenatal development from gastrula to birth.
1972 *Nature* **626**, 1084–1093 (2024).
1973 82. Prummel, K. D., Nieuwenhuize, S. & Mosimann, C. The lateral plate mesoderm. *Development*

1974 **147**, (2020).
1975 83. Prummel, K. D. *et al.* Hand2 delineates mesothelium progenitors and is reactivated in
1976 mesothelioma. *Nature Communications* **13**, 1677 (2022).
1977 84. Segal, J. M. *et al.* Single cell analysis of human foetal liver captures the transcriptional profile of
1978 hepatobiliary hybrid progenitors. *Nat Commun* **10**, 3350 (2019).
1979 85. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature*
1980 **597**, 250–255 (2021).
1981 86. Knight-Schrijver, V. R. *et al.* A single-cell comparison of adult and fetal human epicardium
1982 defines the age-associated changes in epicardial activity. *Nat Cardiovasc Res* **1**, 1215–1229
1983 (2022).
1984 87. Naiche, L. A., Arora, R., Kania, A., Lewandoski, M. & Papaioannou, V. E. Identity and fate of
1985 Tbx4-expressing cells reveal developmental cell fate decisions in the allantois, limb, and external
1986 genitalia. *Dev Dyn* **240**, 2290–2300 (2011).
1987 88. Martik, M. L. & Bronner, M. E. Riding the crest to get a head: neural crest evolution in
1988 vertebrates. *Nature Reviews Neuroscience* **22**, 616–626 (2021).
1989 89. Lee, J. *et al.* Hair-bearing human skin generated entirely from pluripotent stem cells. *Nature* **582**,
1990 399–404 (2020).
1991 90. Merzdorf, C. S. Emerging roles for zic genes in early development. *Dev Dyn* **236**, 922–940
1992 (2007).
1993 91. Inoue, T., Ogawa, M., Mikoshiba, K. & Aruga, J. Zic deficiency in the cortical marginal zone
1994 and meninges results in cortical lamination defects resembling those in type II lissencephaly. *J*
1995 *Neurosci* **28**, 4712–4725 (2008).
1996 92. Gomez-Salinero, J. M., Redmond, D. & Rafii, S. Microenvironmental determinants of
1997 endothelial cell heterogeneity. *Nature Reviews Molecular Cell Biology* **26**, 476–495 (2025).
1998 93. Stone, O. A., Zhou, B., Red-Horse, K. & Stainier, D. Y. R. Endothelial ontogeny and the
1999 establishment of vascular heterogeneity. *Bioessays* **43**, e2100036 (2021).
2000 94. Barnett, S. N. *et al.* An organotypic atlas of human vascular cells. *Nature Medicine* **30**, 3468–
2001 3481 (2024).

2002 95. Kalucka, J. *et al.* Single-cell transcriptome atlas of Murine endothelial cells. *Cell* **180**, 764–
2003 779.e20 (2020).
2004 96. Paik, D. T. *et al.* Single-Cell RNA Sequencing Unveils Unique Transcriptomic Signatures of
2005 Organ-Specific Endothelial Cells. *Circulation* **142**, 1848–1862 (2020).
2006 97. Chen, J. *et al.* Molecular and Spatial Signatures of Mouse Embryonic Endothelial Cells at
2007 Single-Cell Resolution. *Circ Res* **134**, 529–546 (2024).
2008 98. Nolan, D. J. *et al.* Molecular signatures of tissue-specific microvascular endothelial cell
2009 heterogeneity in organ maintenance and regeneration. *Dev Cell* **26**, 204–219 (2013).
2010 99. Wälchli, T. *et al.* Single-cell atlas of the human brain vasculature across development, adulthood
2011 and disease. *Nature* **632**, 603–613 (2024).
2012 100. McCracken, I. R. *et al.* Mapping the developing human cardiac endothelium at single-cell
2013 resolution identifies MECOM as a regulator of arteriovenous gene expression. *Cardiovasc Res*
2014 **118**, 2960–2972 (2022).
2015 101. Jeong, H.-W. *et al.* Single-cell transcriptomics reveals functionally specialized vascular
2016 endothelium in brain. *Elife* **11**, (2022).
2017 102. Tărlungeanu, D. C. *et al.* Impaired Amino Acid Transport at the Blood Brain Barrier Is a Cause
2018 of Autism Spectrum Disorder. *Cell* **167**, 1481–1494.e18 (2016).
2019 103. Kruzynska-Frejtag, A., Machnicki, M., Rogers, R., Markwald, R. R. & Conway, S. J. Periostin
2020 (an osteoblast-specific factor) is expressed within the embryonic mouse heart during valve
2021 formation. *Mech Dev* **103**, 183–188 (2001).
2022 104. Li, Z.-W. *et al.* Oit3, a promising hallmark gene for targeting liver sinusoidal endothelial cells.
2023 *Signal Transduct Target Ther* **8**, 344 (2023).
2024 105. Oliver, G., Kipnis, J., Randolph, G. J. & Harvey, N. L. The Lymphatic Vasculature in the 21
2025 Century: Novel Functional Roles in Homeostasis and Disease. *Cell* **182**, 270–296 (2020).
2026 106. Petrova, T. V. & Koh, G. Y. Biological functions of lymphatic vessels. *Science* **369**, (2020).
2027 107. Stritt, S., Koltowska, K. & Mäkinen, T. Homeostatic maintenance of the lymphatic vasculature.
2028 *Trends Mol. Med.* **27**, 955–970 (2021).
2029 108. Jafree, D. J., Long, D. A., Scambler, P. J. & Ruhrberg, C. Mechanisms and cell lineages in

2030 lymphatic vascular development. *Angiogenesis* **24**, 271–288 (2021).
2031 109. Takeda, A. *et al.* Single-Cell Survey of Human Lymphatics Unveils Marked Endothelial Cell
2032 Heterogeneity and Mechanisms of Homing for Neutrophils. *Immunity* **51**, 561–572.e5 (2019).
2033 110. Bauer, A. *et al.* Transcriptomics- and 3D imaging-based characterization of the lymphatic
2034 vasculature in human skin. *J Exp Med* **223**, (2026).
2035 111. Jafree, D. J. *et al.* Organ-specific features of human kidney lymphatics are disrupted in chronic
2036 transplant rejection. *J. Clin. Invest.* **135**, (2025).
2037 112. Lupu, I.-E. *et al.* Direct specification of lymphatic endothelium from mesenchymal progenitors.
2038 *Nature Cardiovascular Research* **4**, 45–63 (2025).
2039 113. Grimm, L. *et al.* Single-cell analysis of lymphatic endothelial cell fate specification and
2040 differentiation during zebrafish development. *EMBO J* **42**, e112590 (2023).
2041 114. Yamaguchi, S. *et al.* The development of early human lymphatic vessels as characterized by
2042 lymphatic endothelial markers. *EMBO J* **43**, 868–885 (2024).
2043 115. Jafree, D. J. *et al.* Spatiotemporal dynamics and heterogeneity of renal lymphatics in mammalian
2044 development and cystic kidney disease. *Elife* **8**, (2019).
2045 116. Travisano, S. I. *et al.* Single-nuclei multiomic analyses identify human cardiac lymphatic
2046 endothelial cells associated with coronary arteries in the epicardium. *Cell Rep.* **42**, 113106
2047 (2023).
2048 117. Scallan, J. P. *et al.* Foxo1 deletion promotes the growth of new lymphatic valves. *J Clin Invest*
2049 **131**, (2021).
2050 118. Niimi, K., Nakae, J., Inagaki, S. & Furuyama, T. FOXO1 represses lymphatic valve formation
2051 and maintenance via PRDM1. *Cell Rep* **37**, 110048 (2021).
2052 119. Niimi, K. *et al.* FOXO1 regulates developmental lymphangiogenesis by upregulating CXCR4 in
2053 the mouse-tail dermis. *Development* **147**, (2020).
2054 120. Mackie, D. I. *et al.* h mutation causes autosomal recessive nonimmune hydrops fetalis with
2055 lymphatic dysplasia. *J Exp Med* **215**, 2339–2353 (2018).
2056 121. Fritz-Six, K. L., Dunworth, W. P., Li, M. & Caron, K. M. Adrenomedullin signaling is necessary
2057 for murine lymphatic vascular development. *J Clin Invest* **118**, 40–50 (2008).

2058 122. Vaahtomeri, K. *et al.* Locally Triggered Release of the Chemokine CCL21 Promotes Dendritic
2059 Cell Transmigration across Lymphatic Endothelia. *Cell Rep* **19**, 902–909 (2017).
2060 123. Bano, F. *et al.* Structure and unusual binding mechanism of the hyaluronan receptor LYVE-1
2061 mediating leucocyte entry to lymphatics. *Nature Communications* **16**, 2754 (2025).
2062 124. Hernández Vásquez, M. N. *et al.* Transcription factor FOXP2 is a flow-induced regulator of
2063 collecting lymphatic vessels. *EMBO J* **40**, e107192 (2021).
2064 125. Norrmén, C. *et al.* FOXC2 controls formation and maturation of lymphatic collecting vessels
2065 through cooperation with NFATc1. *J. Cell Biol.* **185**, 439–457 (2009).
2066 126. Ortsäter, H., Hernández-Vásquez, M. N., Ulvmar, M. H., Gow, A. & Mäkinen, T. An inducible
2067 Cldn11-CreER mouse line for selective targeting of lymphatic valves. *Genesis* **59**, e23439
2068 (2021).
2069 127. Yin, X. *et al.* Compartmentalized ocular lymphatic system mediates eye-brain immunity. *Nature*
2070 **628**, 204–211 (2024).
2071 128. Doni, A. *et al.* PTX3 is expressed in terminal lymphatics and shapes their organization and
2072 function. *Front Immunol* **15**, 1426869 (2024).
2073 129. Petkova, M. *et al.* Immune-interacting lymphatic endothelial subtype at capillary terminals drives
2074 lymphatic malformation. *J Exp Med* **220**, (2023).
2075 130. Xiang, M. *et al.* A Single-Cell Transcriptional Roadmap of the Mouse and Human Lymph Node
2076 Lymphatic Vasculature. *Front Cardiovasc Med* **7**, 52 (2020).
2077 131. Klotz, L. *et al.* Cardiac lymphatics are heterogeneous in origin and respond to injury. *Nature*
2078 **522**, 62–67 (2015).
2079 132. Maruyama, K., Miyagawa-Tomita, S., Mizukami, K., Matsuzaki, F. & Kurihara, H. Isl1-
2080 expressing non-venous cell lineage contributes to cardiac lymphatic vessel development. *Dev*
2081 *Biol* **452**, 134–143 (2019).
2082 133. Maruyama, K. *et al.* The cardiopharyngeal mesoderm contributes to lymphatic vessel
2083 development in mouse. *Elife* **11**, (2022).
2084 134. Lioux, G. *et al.* A Second Heart Field-Derived Vasculogenic Niche Contributes to Cardiac
2085 Lymphatics. *Dev Cell* **52**, 350–363.e6 (2020).

2086 135. Wu, Y. *et al.* Organogenesis and distribution of the ocular lymphatic vessels in the anterior eye.
2087 *JCI Insight* **5**, (2020).
2088 136. Izen, R. M., Yamazaki, T., Nishinaka-Arai, Y., Hong, Y.-K. & Mukouyama, Y.-S. Postnatal
2089 development of lymphatic vasculature in the brain meninges. *Dev. Dyn.* **247**, 741–753 (2018).
2090 137. Antila, S. *et al.* Development and plasticity of meningeal lymphatic vessels. *J Exp Med* **214**,
2091 3645–3667 (2017).
2092 138. Albayram, M. S. *et al.* Non-invasive MR imaging of human brain lymphatic networks with
2093 connections to cervical lymph nodes. *Nat Commun* **13**, 203 (2022).
2094 139. Louveau, A. *et al.* Structural and functional features of central nervous system lymphatic vessels.
2095 *Nature* **523**, 337–341 (2015).
2096 140. Absinta, M. *et al.* Human and nonhuman primate meninges harbor lymphatic vessels that can be
2097 visualized noninvasively by MRI. *Elife* **6**, (2017).
2098 141. Jacob, L. *et al.* Conserved meningeal lymphatic drainage circuits in mice and humans. *J Exp Med*
2099 **219**, (2022).
2100 142. Mezey, É. *et al.* An immunohistochemical study of lymphatic elements in the human brain. *Proc*
2101 *Natl Acad Sci U S A* **118**, (2021).
2102 143. Martin-Almedina, S., Mortimer, P. S. & Ostergaard, P. Development and physiological functions
2103 of the lymphatic system: insights from human genetic studies of primary lymphedema. *Physiol*
2104 *Rev* **101**, 1809–1871 (2021).
2105 144. Gordon, K. *et al.* Update and audit of the St George's classification algorithm of primary
2106 lymphatic anomalies: a clinical and molecular approach to diagnosis. *J Med Genet* **57**, 653–659
2107 (2020).
2108 145. Fotiou, E. *et al.* Novel mutations in PIEZO1 cause an autosomal recessive generalized lymphatic
2109 dysplasia with non-immune hydrops fetalis. *Nat Commun* **6**, 8085 (2015).
2110 146. Lukacs, V. *et al.* Impaired PIEZO1 function in patients with a novel autosomal recessive
2111 congenital lymphatic dysplasia. *Nat Commun* **6**, 8329 (2015).
2112 147. Choi, D. *et al.* Piezo1 incorporates mechanical force signals into the genetic program that
2113 governs lymphatic valve development and maintenance. *JCI Insight* **4**, (2019).

2114 148. Nonomura, K. *et al.* Mechanically activated ion channel PIEZO1 is required for lymphatic valve
2115 formation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 12817–12822 (2018).
2116 149. Soldatov, R. *et al.* Spatiotemporal structure of cell fate decisions in murine neural crest. *Science*
2117 **364**, (2019).
2118 150. Thiery, A. P. *et al.* scRNA-sequencing in chick suggests a probabilistic model for cell fate
2119 allocation at the neural plate border. *Elife* **12**, (2023).
2120 151. Lu, T. *et al.* Decoding transcriptional identity in developing human sensory neurons and
2121 organoid modeling. *Cell* **187**, 7374–7393.e28 (2024).
2122 152. Kameneva, P. *et al.* Single-cell transcriptomics of human embryos identifies multiple
2123 sympathoblast lineages with potential implications for neuroblastoma origin. *Nat Genet* **53**, 694–
2124 706 (2021).
2125 153. Aivazidis, A. *et al.* Cell2fate infers RNA velocity modules to improve cell fate prediction. *Nat.*
2126 *Methods* **22**, 698–707 (2025).
2127 154. Kastriti, M. E. *et al.* Schwann cell precursors represent a neural crest-like state with biased
2128 multipotency. *EMBO J* **41**, e108780 (2022).
2129 155. Simões-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe.
2130 *Development* **142**, 242–257 (2015).
2131 156. Laddach, A. *et al.* A branching model of lineage differentiation underpinning the neurogenic
2132 potential of enteric glia. *Nat Commun* **14**, 5904 (2023).
2133 157. Brugmann, S. A., Pandur, P. D., Kenyon, K. L., Pignoni, F. & Moody, S. A. Six1 promotes a
2134 placodal fate within the lateral neurogenic ectoderm by functioning as both a transcriptional
2135 activator and repressor. *Development* **131**, 5871–5881 (2004).
2136 158. Edens, B. M. & Bronner, M. E. Making sense of vertebrate senses from a neural crest and cranial
2137 placode evo-devo perspective. *Trends Neurosci* **48**, 213–226 (2025).
2138 159. Rayon, T., Maizels, R. J., Barrington, C. & Briscoe, J. Single-cell transcriptome profiling of the
2139 human developing spinal cord reveals a conserved genetic programme with human-specific
2140 features. *Development* **148**, (2021).
2141 160. Murtazina, A. & Adameyko, I. The peripheral nervous system. *Development* **150**, (2023).

2142 161. Kupari, J., Häring, M., Agirre, E., Castelo-Branco, G. & Ernfors, P. An Atlas of Vagal Sensory
2143 Neurons and Their Molecular Specialization. *Cell Rep* **27**, 2508–2523.e4 (2019).
2144 162. Bina, P., Hines, M. A., Sanyal, J. & Taneyhill, L. A. Neurogenin 2 and Neuronal Differentiation
2145 1 Control Proper Development of the Chick Trigeminal Ganglion and Its Nerve Branches. *J Dev*
2146 *Biol* **11**, (2023).
2147 163. Fawkner-Corbett, D. *et al.* Spatiotemporal analysis of human intestinal development at single-
2148 cell resolution. *Cell* **184**, 810–826.e23 (2021).
2149 164. Carmeliet, P. & Tessier-Lavigne, M. Common mechanisms of nerve and blood vessel wiring.
2150 *Nature* **436**, 193–200 (2005).
2151 165. Marmigère, F. & Ernfors, P. Specification and connectivity of neuronal subtypes in the sensory
2152 lineage. *Nat Rev Neurosci* **8**, 114–127 (2007).
2153 166. Lallemend, F. & Ernfors, P. Molecular interactions underlying the specification of sensory
2154 neurons. *Trends Neurosci* **35**, 373–381 (2012).
2155 167. Le, V.-H. *et al.* CXCL12 promotes the crossing of retinal ganglion cell axons at the optic chiasm.
2156 *Development* **151**, (2024).
2157 168. Williams, E. J., Furness, J., Walsh, F. S. & Doherty, P. Activation of the FGF receptor underlies
2158 neurite outgrowth stimulated by L1, N-CAM, and N-cadherin. *Neuron* **13**, 583–594 (1994).
2159 169. Smeyne, R. J. *et al.* Severe sensory and sympathetic neuropathies in mice carrying a disrupted
2160 Trk/NGF receptor gene. *Nature* **368**, 246–249 (1994).
2161 170. Crowley, C. *et al.* Mice lacking nerve growth factor display perinatal loss of sensory and
2162 sympathetic neurons yet develop basal forebrain cholinergic neurons. *Cell* **76**, 1001–1011
2163 (1994).
2164 171. Fariñas, I., Jones, K. R., Backus, C., Wang, X. Y. & Reichardt, L. F. Severe sensory and
2165 sympathetic deficits in mice lacking neurotrophin-3. *Nature* **369**, 658–661 (1994).
2166 172. Marcu, R. *et al.* Human Organ-Specific Endothelial Cell Heterogeneity. *iScience* **4**, 20–35
2167 (2018).
2168 173. Lin, L. *et al.* A single-cell time-series atlas of endothelial cell embryonic development. *Cell*
2169 (2026) doi:10.1016/j.cell.2026.01.002.

2170 174. Aragon, J. W. *et al.* Emergence of endothelial subtypes and role of cell cycle control in arterial-
2171 venous specification during embryonic vascular development. *Cell Rep* **44**, 116368 (2025).
2172 175. Jerafi-Vider, A. *et al.* VEGFC/FLT4-induced cell-cycle arrest mediates sprouting and
2173 differentiation of venous and lymphatic endothelial cells. *Cell Rep* **35**, 109255 (2021).
2174 176. Saint-Jeannet, J.-P. & Moody, S. A. Establishing the pre-placodal region and breaking it into
2175 placodes with distinct identities. *Dev Biol* **389**, 13–27 (2014).
2176 177. Bahadir, C. D. *et al.* Artificial intelligence applications in histopathology. *Nature Reviews*
2177 *Electrical Engineering* **1**, 93–108 (2024).
2178 178. Miller, A. J. *et al.* In Vitro and In Vivo Development of the Human Airway at Single-Cell
2179 Resolution. *Dev. Cell* **53**, 117–128.e6 (2020).
2180 179. Sridhar, A. *et al.* Single-cell transcriptomic comparison of human fetal retina, hPSC-derived
2181 retinal organoids, and long-term retinal cultures. *Cell Rep.* **30**, 1644–1659.e4 (2020).
2182 180. Yu, Q. *et al.* Charting human development using a multi-endodermal organ atlas and organoid
2183 models. *Cell* **184**, 3281–3298.e22 (2021).
2184 181. Inecik, K., Erken, E. & Theis, F. J. IDTrack: Time- and Namespace-Aware Identifier
2185 Harmonization for Reproducible Workflows. *bioRxiv* 2026.02.05.703984 (2026)
2186 doi:10.64898/2026.02.05.703984.
2187 182. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for
2188 single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
2189 183. De Donno, C. *et al.* Population-level integration of single-cell datasets enables multi-scale
2190 analysis across samples. *Nature Methods* **20**, 1683–1692 (2023).
2191 184. Inecik, K., Kara, A., Rose, A., Haniffa, M. & Theis, F. J. TarDis: Achieving Robust and
2192 Structured Disentanglement of Multiple Covariates. *bioRxiv* 2024.06.20.599903 (2024)
2193 doi:10.1101/2024.06.20.599903.
2194 185. Garcia-Alonso, L. *et al.* Single-cell roadmap of human gonadal development. *Nature* **607**, 540–
2195 547 (2022).
2196 186. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data
2197 analysis. *Genome Biol.* **19**, 15 (2018).

2198 187. Dimitrov, D. *et al.* LIANA+ provides an all-in-one framework for cell–cell communication
2199 inference. *Nature Cell Biology* **26**, 1613–1622 (2024).
2200 188. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference
2201 from single-cell RNA-Seq data. *Nat. Commun.* **13**, 3224 (2022).
2202 189. Badia-I-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological
2203 activities from omics data. *Bioinform Adv* **2**, vbac016 (2022).
2204 190. Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis
2205 of single-cell omics data. *Nat Methods* **21**, 217–227 (2024).
2206 191. Feng, J., Liu, T. & Zhang, Y. Using MACS to Identify Peaks from ChIP-Seq Data. *Current*
2207 *protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **CHAPTER**,
2208 Unit2.14 (2011).
2209 192. Yates, T. M. *et al.* Curating genomic disease-gene relationships with Gene2Phenotype (G2P).
2210 *Genome Medicine* **16**, 1–8 (2024).
2211 193. Foreman, J. *et al.* DECIPHER: Improving genetic diagnosis through dynamic integration of
2212 genomic and clinical data. *Annu. Rev. Genomics Hum. Genet.* **24**, 151–176 (2023).