

Gene expression dynamics of human and mouse craniofacial development at the single-cell level

Received: 23 January 2025

Accepted: 19 February 2026

Published online: 09 March 2026

 Check for updates

Nagham Khouri-Farah ^{1,7}, Alexandra Manchel ^{2,7},
Emma Wentworth Winchester¹, Brian M. Schilder ^{3,4}, Kelsey Robinson⁵,
Sarah W. Curtis⁵, Nathan G. Skene ^{3,4}, Elizabeth J. Leslie-Clarkson ⁵ &
Justin Cotney ^{2,6} 

Craniofacial development is a complex process that involves the specification of diverse and transient cell types. However, our understanding of these processes and the cell-types present during human craniofacial development remains limited. We address this gap in knowledge through single-nucleus RNA sequencing of human craniofacial development spanning 4 to 8 post-conception weeks. This resource identifies multiple subtypes of mesenchyme, epithelium, and cranial neural crest, among other functionally distinct cell types. Extensive comparisons to single-nucleus gene expression and spatial transcriptomics of comparable mouse developmental stages reveal functional conservation of most cell types and identification of anatomically distinct cell subtypes. We find distinct contributions of cellular subtypes to normal facial morphology and common risk factors for orofacial clefts. Additionally, we find specific ectodermal and epithelial subtypes whose gene expression networks are most significantly affected by rare de novo protein-altering variants from orofacial cleft patients. Together our data provide an extensive resource for understanding human craniofacial development at the cellular level.

Craniofacial development orchestrates the formation of the human face through the interplay of multiple cell types, including mesenchyme, ectoderm, endothelium, and cranial neural crest cells (CNCCs), which differentiate into a diverse array of tissues such as bone, cartilage, muscle, skin, and vasculature^{1–3}. Together, these cells and tissues perform essential functions of the face like respiration, mastication, communication, and sensory perception^{4–7}. Disruptions to craniofacial developmental processes rank amongst the most common causes of human congenital anomalies, with orofacial clefts representing a significant portion of global birth defects^{8–11}. Thus, there exists significant

need to understand the molecular, genetic, and cellular mechanisms underlying human craniofacial development.

Studies utilizing model organisms, particularly mice, have offered key insights into craniofacial development and abnormalities^{12–16}. However, significant differences exist between mouse and human craniofacial development, including variations in timing, cellular contributions, and gene regulatory networks^{1,13,17–22}. Furthermore, human craniofacial features exhibit evolutionary adaptations that distinguish them from other mammals and primates, underscoring the necessity for human-specific studies^{5,13,21,22}.

¹Graduate Program in Genetics and Developmental Biology, UConn Health, Farmington, CT, USA. ²Department of Surgery, Children's Hospital of Philadelphia, Philadelphia, PA, USA. ³Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, UK. ⁴UK Dementia Research Institute at Imperial College London, London, UK. ⁵Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA. ⁶Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ⁷These authors contributed equally: Nagham Khouri-Farah, Alexandra Manchel.

 e-mail: cotneyj@chop.edu

Advances in single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq, respectively) technologies have enabled detailed characterization of cellular diversity and gene expression during development^{15,23–27}. These tools are particularly valuable for resolving the dynamics of rare or transient cell populations, such as CNCCs, that play critical roles in craniofacial formation¹³. While previous efforts have developed single-cell atlases for murine craniofacial tissues, corresponding human datasets are limited^{15,23}. Few studies have examined bulk gene expression during human craniofacial development^{13,14,28–31}, with only two datasets sampled during the embryonic period of human development¹³. While mapping of human genetics findings to mouse craniofacial cell types has indicated potential disease-causing subtypes³², limited replicates and differences between human and mouse craniofacial development preclude confident interpretation.

To address these gaps in knowledge, we constructed a time-resolved gene expression atlas of human craniofacial development by collecting craniofacial tissues from 24 individual human embryos encompassing six key time points from 4 to 8 post-conception weeks^{33,34}. We profiled over 95,000 nuclei and identified eight major cell types, including mesenchyme, ectoderm, endothelium, muscle progenitors, *SOX10*+ CNCCs, and *SOX2*+ early progenitors. Comparative analysis with newly generated and previously published²³ murine craniofacial datasets highlighted significant conservation of major cell types and their gene expression programs. Integration with human spatial transcriptomics further validated the localization of these subtypes within the developing human face.

Additionally, this study explores the genetic factors shaping human craniofacial features. By integrating genome-wide association studies (GWAS) with our atlas, we identified cell type-specific enrichments for genetic variants associated with normal facial variation. We found that specific subtypes of ectoderm and mesenchyme contribute to different aspects of facial appearance and shape. We also examined rare variants associated with congenital craniofacial disorders and found that *de novo* protein-damaging variants identified in orofacial clefting trios were enriched in genes that specify distinct ectodermal and epithelial subtypes.

This comprehensive atlas provides a high-resolution view of the cellular and molecular landscape of human craniofacial development, integrating gene expression, spatial mapping, and evolutionary genomics. Our work enhances our understanding of human craniofacial biology and establishes a framework for future studies aimed at uncovering therapeutic targets and evolutionary insights into craniofacial biology. These data can be explored through an interactive web application accessible via https://cotneyshiny.research.chop.edu/shiny-apps/craniofacial_all_snRNA_v2/ and at the Chan-Zuckerberg CellXGene Discover resource³⁵.

Results

Single nuclei gene-expression atlas of the developing human face

To characterize the cellular landscape of human craniofacial development we performed snRNA-seq analysis on the entire craniofacial prominence of 24 individual human embryos across 6 distinct time points, encompassing major milestones of human craniofacial development from 4 to 8 post conception weeks (Fig. 1a). Following multiple levels of quality control and filtering (“Methods” and Supplementary Fig. 1) we obtained gene expression profiles from 95,892 nuclei with an average of 3996 nuclei from each sample (Supplementary Fig. 2A) and a median of 6890 counts from 2988 genes per nucleus (Supplementary Fig. 2B, C). To compare to our previously published bulk gene expression analysis, we combined data from all nuclei of a specific stage into pseudo-bulk gene expression profiles. This revealed the greatest similarity between samples separated by the least developmental time (Supplementary Fig. 3). Principal component and

differential expression analyses of these pseudo-bulk profiles revealed very similar results to those previously obtained from bulk tissues¹³. Specifically, samples were ordered by developmental stage along the first principal component (Fig. 1b and Supplementary Fig. 4A) and the greatest number of differentially expressed genes were observed between the earliest and latest comparable timepoints (CS13 vs CS17) (Supplementary Fig. 4B, C).

Clustering and projection of these single nuclei expression profiles in two dimensions using Uniform Manifold Approximation and Projection (UMAP)³⁶ revealed eight distinct populations (Fig. 1c). The clusters were contributed to by samples from each of the stages in very similar proportions (Fig. 1d, e). To characterize the eight major clusters, we first examined expression of four genes defined by Li et al.¹², *ALX1*, *EPCAM*, *CDH5*, and *FCERG1* as markers of mesenchyme, ectoderm, endothelium, and immune cells, respectively (Fig. 2a). *ALX1* was most strongly expressed in cluster 0 (mesenchyme, a tissue with dual origin from the mesoderm and cranial neural crest during development), *EPCAM* in cluster 1 (ectoderm), *CDH5* in cluster 4 (endothelium), and *FCERG1* in cluster 6 (immune). We also examined markers of erythrocytes, including *SPTA1*, *ALAS2*, *RHAG*, and *HEGMN*, which were strongly enriched in cluster 3 (Supplementary Fig. 5). Clusters 2, 5, and 7 did not show signal for any of these genes.

To characterize the remaining clusters, we identified the top 5 genes that were most strongly differentially expressed between the clusters (Fig. 2b and Supplementary Data 1). In agreement with the initial cluster identities suggested by the markers of Li et al.¹², cluster 0 (mesenchyme) was marked by *SATB2*^{37,38} and *RSPO2*^{39,40}, cluster 1 (ectoderm) was identified by *GRHL2*⁴¹ and *TP63*, cluster 3 (erythrocytes) was marked by *SPTA1*^{42,43} and *RHAG*^{44,45}, cluster 4 (endothelium) showed highly biased expression for *FLT1*⁴⁶ and *MYCT1*⁴⁷, and cluster 6 (immune) was marked by *MPEG1*⁴⁸, *CD86*⁴⁹, and *CD163*^{49–52}. Cluster 2 showed highly biased expression of *MYOG*, *MYH3*, and *MYBPH*, all genes related to muscle specification and function^{53–55}, while cluster 5 was marked by *FOXD3*, a developmental transcription factor which has been linked to pluripotency maintenance in stem cells and specification of neural crest in multiple species^{56–59}. This cluster was also marked by high expression of *SOX10* but not *SOX2* (Fig. 2a). Thus, we labeled this cluster as *SOX2*–/*SOX10*+ putative neural crest. Cluster 7 showed highly biased expression of *SOX2*^{60,61}, *PAX6*⁶², and *POU3F2*⁶³, which are all markers of early progenitor cells (Fig. 2a). However, these cells did not express *SOX10* and thus were labeled as *SOX2*+/*SOX10*– early progenitors. We inferred RNA velocity and transition states to identify trajectories in the data across the newly identified main cell types. We found that the trajectories emanate from the *SOX2*+/*SOX10*– population, confirming that these cells likely represent an early progenitor population (Fig. 2c).

To determine biological and disease relevance of each of these clusters, we analyzed the top 100 marker genes from each cluster for gene ontology, pathway, and disease enrichments. The genes that defined the mesenchyme were enriched for processes related to extracellular matrix development and skeletal, cartilage, and roof of mouth development as well as diseases including malar flattening and isolated cleft palate (Fig. 2d, Supplementary Fig. 6, and Supplementary Data 2). Genes most strongly expressed in ectoderm were enriched for processes related to tight junctions, cell adhesion, and the plasma membrane and diseases including Epidermolysis Bullosa, Keratoderma, and Palate fistula (Fig. 2d, Supplementary Fig. 6, and Supplementary Data 2). The marker genes identified in the *SOX2*–/*SOX10*+ plausible cranial neural crest cell (CNCC) population were enriched for processes related to nervous system development, pathways related to *EGR2* and *SOX10*-mediated initiation of Schwann cell myelination and diseases related to neuropathies, primary brain tumors, and neural crest tumors (Fig. 2d, Supplementary Fig. 6, and Supplementary Data 2). Marker genes for the erythrocyte, immune, endothelial,

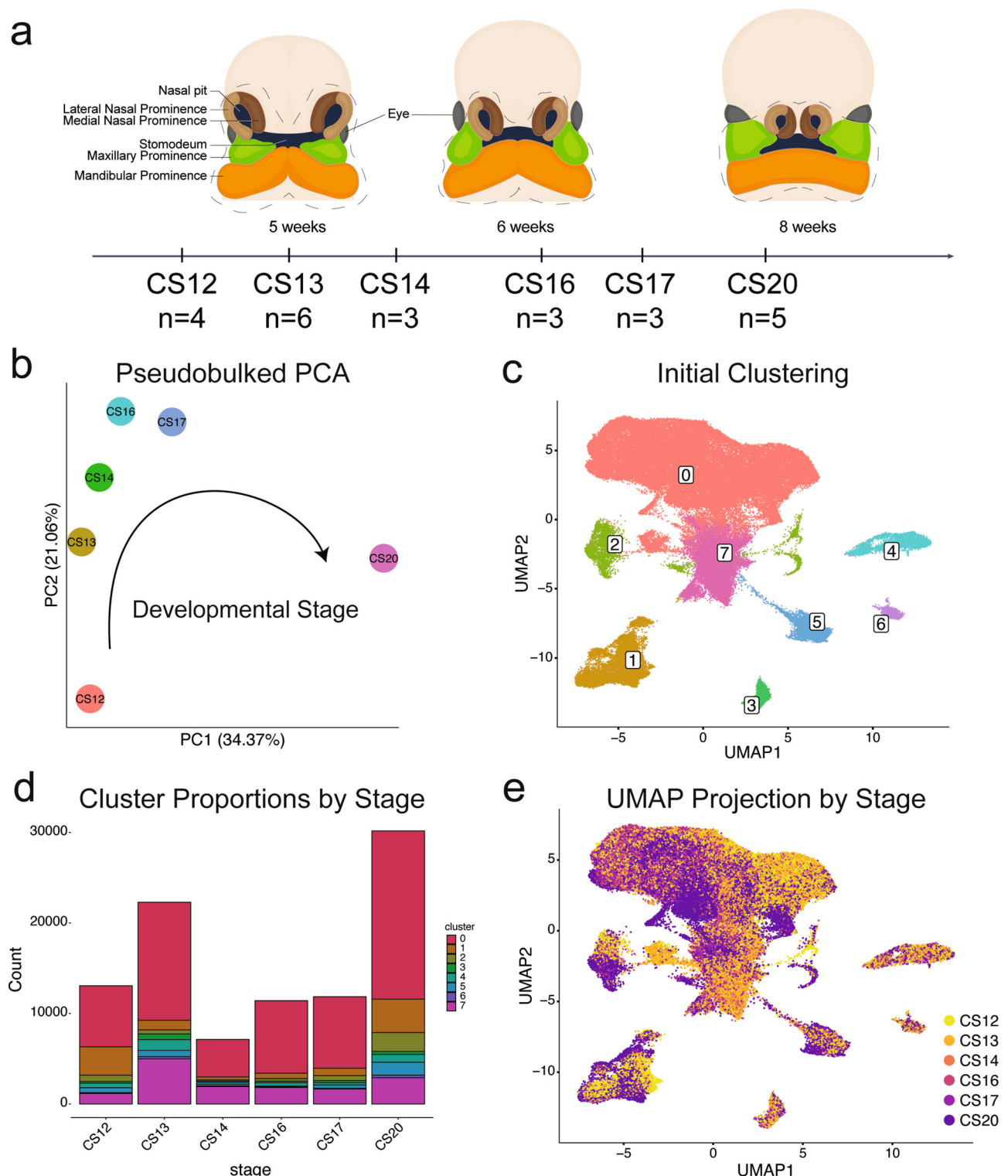


Fig. 1 | Generation of single nucleus gene expression atlas of human craniofacial development. **a** Dashed regions indicate approximate anatomical regions of the developing craniofacial region, with the exclusion of the eyes from 4 to 8 weeks post conception, that were collected by HDBR for processing in this study. The same tissue source and collection methods were previously described in Wilderman et al.²⁸ and Rajderkar et al.¹⁴. Individual Carnegie Stages (CS) and replicates at stage are indicated below images. Adapted from Selleri and Rijli 2023²¹.

Orthologous regions were collected in this study from mouse samples ranging from E10.5 to E12.5. **b** Pseudo-bulk gene expression of tissues from each stage displayed in principal component (PC) space based on the first two PCs. Progression of developmental time is indicated along PC1 dimension. **c** UMAP projection and cluster identification of all human craniofacial cells after filtering of neurons. **d** Number of cells obtained at each CS stage for each cluster identified in this figure (c). **e** Distribution of cells from each stage across the UMAP projection.

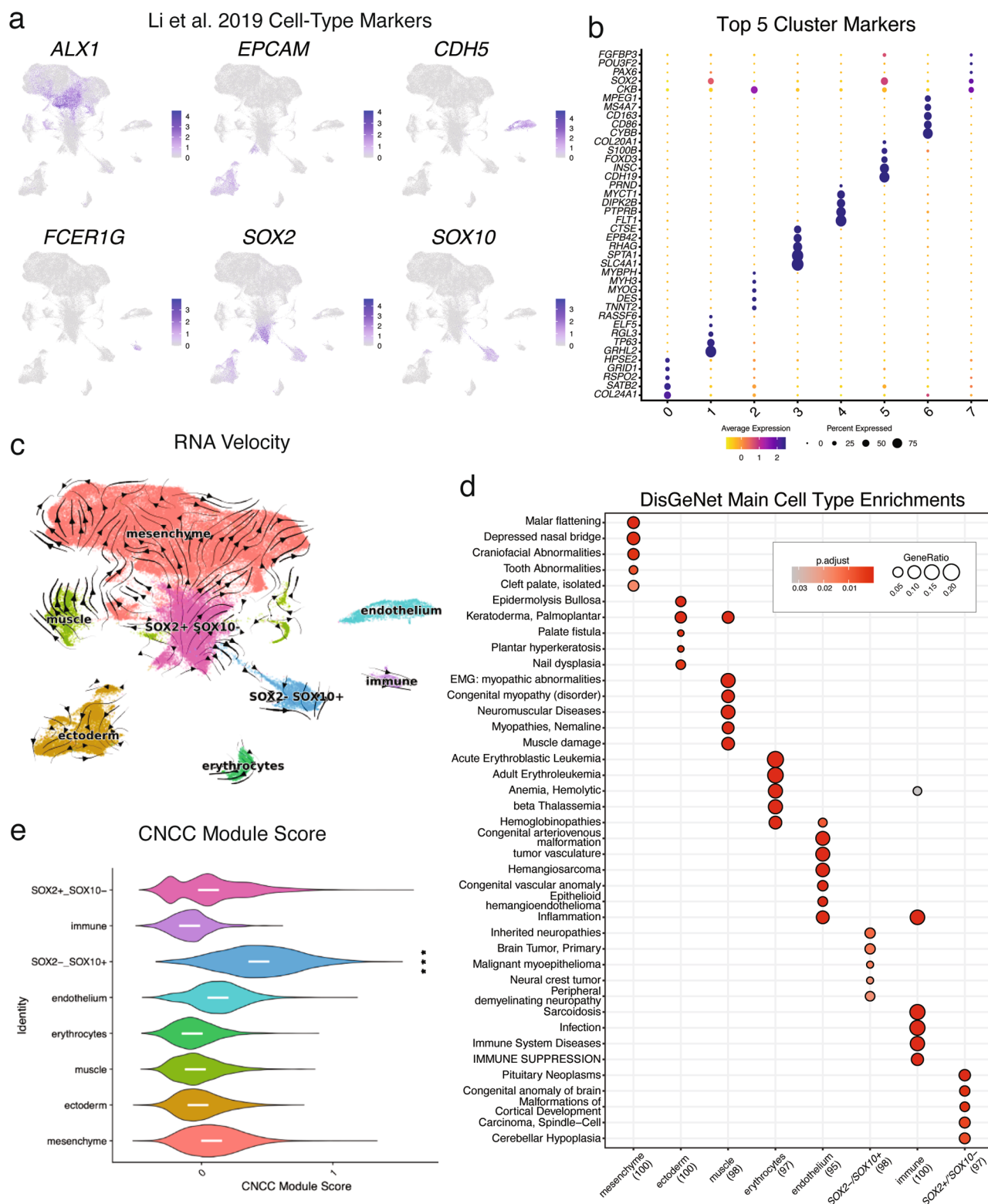


Fig. 2 | Identification of main cell types in the developing human face. **a** Gene expression feature plots for indicated genes across UMAP projection. **b** Average and percent expression for the top 5 marker genes for each main cluster. **c** RNA velocity trajectories displayed on UMAP projection of main cell types. **d** Disease ontology enrichments of categories curated by DisGeNet for each main cell type.

e Violin plot for all cells of a given main cell type based on CNCC module score of curated neural crest genes calculated by Seurat. Significance testing between *SOX2*⁺/*SOX10*⁺ cells vs. other main cell types were calculated using the Wilcox test. Raw *p*-values were Benjamini–Hochberg adjusted using the *p.adjust* function in the stats RStudio package. Padj < 0.001 = ***, Padj < 0.01 = **, Padj < 0.05 = *.

muscle progenitor, and *SOX2*⁺/*SOX10*[−] early progenitor clusters also showed enrichment of relevant cell type-specific processes and disease ontologies (Supplementary Fig. 6, Fig. 2d, and Supplementary Data 2).

Characterization of putative human cranial neural crest

To further characterize the *SOX2*⁺/*SOX10*⁺ presumptive cranial neural crest (CNCC) cluster, we performed additional analyses using known cranial neural crest markers, including *EDNRB*, *ERBB3*, *PAX3*, *SOX10*,

SPPI, *TFAP2B*, and *ZEB2*, all genes well known to be involved in various aspects of neural crest specification and migration⁶⁴. While these genes are biased toward the *SOX2*–/*SOX10*+ putative CNCC cluster, they are not exclusively expressed in cells found in this cluster (Supplementary Fig. 7). Since no single gene was capable of confirming the identity of this population, we generated a module score of genes from regulatory networks identified in cultured human and chimpanzee CNCCs⁶⁵. We found significantly higher expression of this module in the *SOX2*–/*SOX10*+ putative CNCC cell population (Fig. 2e), confirming our initial labeling.

Since few studies have been able to identify populations of CNCCs from primary human tissue^{66–68}, we sought to subcluster this putative CNCC population ($n = 3294$) to better understand the gene expression programs active in these cells. We identified 19 distinct subclusters from this original population, some of which were biased towards early vs. late time points (Fig. 3a, b and Supplementary Fig. 8A–C). We analyzed the genes enriched in each of these subclusters to label based on functionality. For instance, CNCC cluster 16 was enriched for melanocyte-specific gene markers *MITF* and *PMEL* (Supplementary Fig. 8D), while two other clusters (clusters 10 and 14) were significantly enriched for the G2M stage of the cell cycle and labeled as cycling (Supplementary Fig. 8E). Since *SOX10* is also a known marker for neural crest-derived Schwann cells⁶⁹, we generated a Schwann cell module score using marker genes from human fetal development²⁴. We identified five clusters that were enriched for the Schwann cell module score (Supplementary Fig. 8F) and were biased towards the later time points profiled (Supplementary Fig. 8C). To identify additional subclusters, we generated additional module scores from markers of E9.5 mouse neural crest⁷⁰ labelled by *Wnt1*^{Cre}/*R26R*^{Tomato}, which consist of various functionally specific, fate-committed CNCC subtypes. Using this methodology, we identified one cluster of mesenchymal-committed CNCCs and one cluster of sensory-committed CNCCs (Supplementary Fig. 8G). We also identified four more punctate clusters, which we annotated as neural crest like (ncl). The resulting subclusters were not differentially enriched for any of the previously computed module scores and were labeled according to stage-specific bias, i.e., early (early.CNCC) vs. intermediate (int.CNCC) (Fig. 3b and Supplementary Fig. 8H). While *SOX10*, *NR2F1*, and *NR2F2* were observed in all subclusters, *TFAP2B* expression was lower in the intermediate CNCCs and *TFAP2A* was absent from the int.CNCC3 cluster. Interestingly, *SOX9* was specifically expressed in the mesenchymal-committed CNCCs (Fig. 3c).

We identified marker genes for each of the putative CNCC subtypes and found 5,279 genes that were differentially expressed (Supplementary Data 3). We highlighted the top five transcription factor marker genes that distinguished each subtype, including *MEIS2*, *MECOM*, and *PAX3* in early CNCCs; *DACH1* and *BCL11B* in the int.CNCC.1 subcluster; *FLI1* and *ALX1* in mesenchymal-committed CNCCs, and *NEUROG2* in sensory-committed CNCCs (Fig. 3d, Supplementary Fig. 8I, and Supplementary Data 3). Identification of enriched gene ontology categories for each subtype revealed distinct functions (Supplementary Fig. 9 and Supplementary Data 4). Early CNCCs were enriched for early pattern specification and axon guidance, and intermediate CNCCs were enriched for ATP synthesis and oxidative phosphorylation. Presumptive melanocytes were enriched for melanin biosynthetic processes and developmental pigmentation, while Schwann cells were enriched for synapse organization and sympathetic nervous system development. The ncl1 cluster was enriched for several categories shared with early CNCCs and Schwann cells, suggesting this was an early neural crest type with possible downstream fate commitment towards a Schwann cell phenotype. The ncl2 cluster was enriched for processes overlapping with the mesenchymal-committed CNCCs, including hair follicle development and skin development. The ncl3 cluster was enriched for synaptic signaling, while the ncl4 was enriched for gas transport processes. We also

examined the previously generated CNCC module score (Fig. 2e) across subtypes identified from ectoderm and mesenchyme clusters, which we describe in further detail below, and found that all subtypes labeled as *SOX2*–/*SOX10*+ putative CNCC have high module scores, while those from other main cell type clusters are considerably lower (Supplementary Fig. 8J).

To determine if these subclusters have relevant spatial patterns of expression, we examined recently published spatial transcriptomics on two sections of a human CS13 embryo²⁵ (“Methods” and Supplementary Fig. 10). We found expression of marker genes that identified CNCCs versus the other major craniofacial cell types had both broad and restricted patterns of expression. For example, *NR2F1* was broadly expressed across the embryo, whereas *PAX3*, *TFAP2A*, and *TFAP2B* were more regionally restricted to the head and neural tube regions (Fig. 3e and Supplementary Fig. 11A). Genes identified as markers of CNCC subtypes showed unique patterns of expression with *MECOM*, a marker of early.CNCC.1 and *TWIST1*, a marker of mesenchymal CNCCs, expressed in the craniofacial region as well as putative somites. *ALX4* was generally restricted to the head region and putative frontal nasal process region, while *HAND2* was expressed in the putative pharyngeal arch regions, heart, and limb (Supplementary Fig. 11B). We then calculated module scores on these spatial data using the marker genes from each of the CNCC subtypes. We found that each of these marker genes sets were biased toward the neural tube region (Supplementary Fig. 11C), consistent with their initial site of specification.

In summary while the *SOX2*–/*SOX10*+ cluster is clearly enriched for gene expression signatures of CNCCs, not all subclusters are CNCCs. Instead, some subclusters represent more terminally differentiated cell states including Schwann cells and melanocytes. Moreover, since the neural crest have been somewhat elusive to detect in previous single cell atlases it is possible other CNCC states are present in the developing embryo that we are unable to sample here.

Conservation of cell types in craniofacial development

Next, we sought to determine if these subtypes are present in mouse and if similar sets of genes specify all the main cell types we found. To allow direct comparisons to our human data, we generated snRNA-seq data from mouse craniofacial tissues harvested from multiple biological replicates of E10.5 to E12.5, which reflect the major morphological landmarks of the human tissue profiled. We integrated these data with recently published single cell gene expression data from E13.5 and E15.5²³, resulting in 79402 expression profiles (Supplementary Fig. 12, “Methods”). When we clustered these data using approaches identical to the human data, we obtained seven main clusters with remarkably similar cluster ratios and organization in the UMAP projection space (Fig. 4a and Supplementary Data 5). When we examined gene expression of the same major markers profiled in human (Fig. 2a), we readily identified the same major cell types including the muscle and *Sox2*–/*Sox10*+ putative CNCC cell types (Fig. 4b). To determine if these cell types were specified by the same sets of genes, we compared marker gene identities and found significant sharing of marker genes between the orthologous major cell types (Fig. 4c). The highest degree of sharing was found between endothelium, followed by mesenchyme and erythrocytes. While still significant, the lowest degree of marker gene conservation was observed between the *Sox2*–/*Sox10*+ putative CNCC cell type cluster of each species (Fig. 4c). Despite these substantial similarities, we did not recover a separate *Sox2*+ progenitor cluster in mouse craniofacial tissue.

Since the majority of main cell types were functionally conserved between human and mouse, we wondered if we could identify conserved subtypes and infer spatial identities for human subtypes. To address this question, we leveraged the substantial annotation resources that have been recently generated for mouse craniofacial development^{12,14,23}. We first independently subclustered the mouse CNCCs in an identical manner to that of human, resulting in 12 distinct

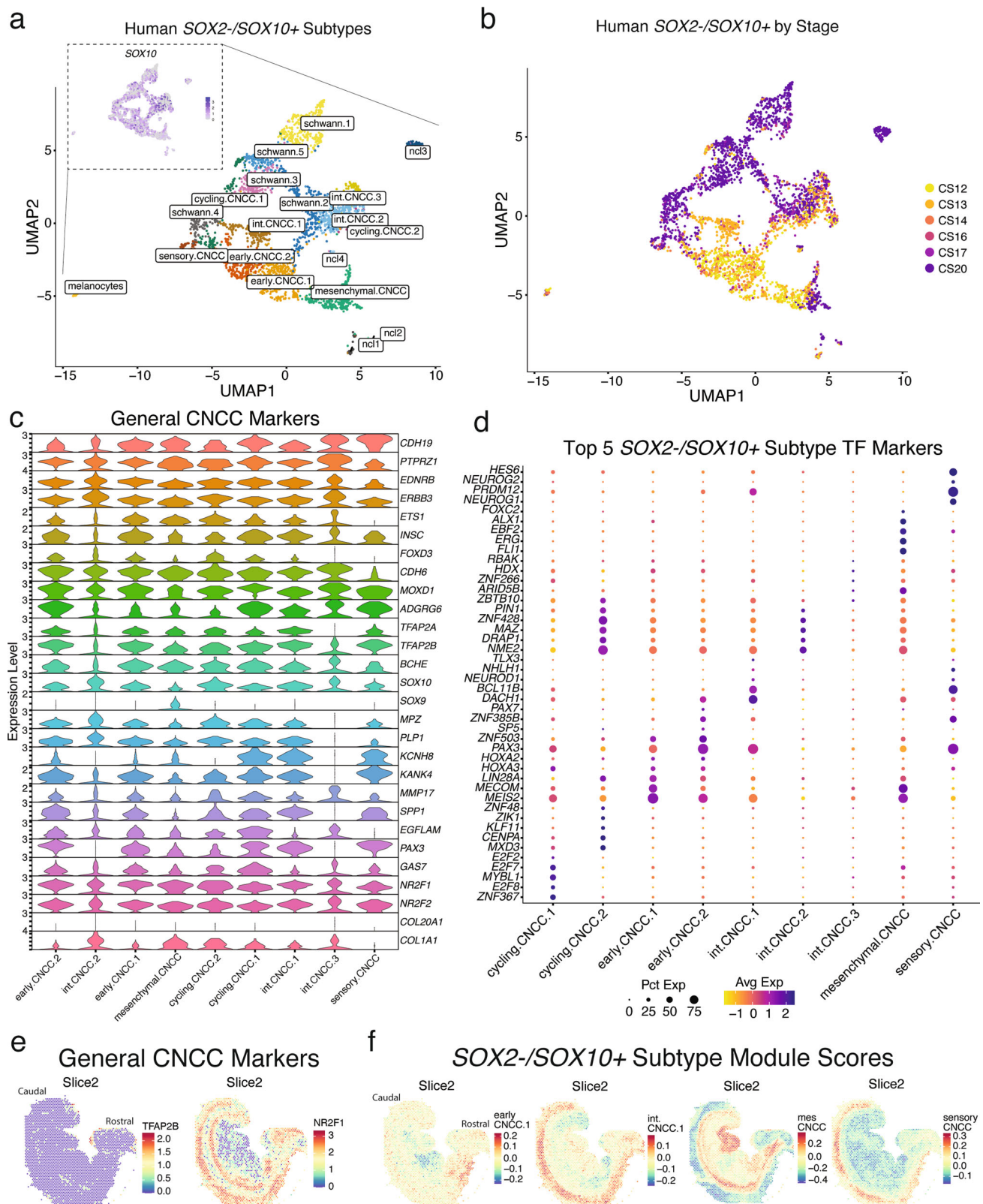


Fig. 3 | Identification of CNCC subtypes in the developing human face. a UMAP projection of subclustered CNCC main cell type. **b** Distribution of cells from each stage across the CNCC subclustered UMAP projection. **c** Violin plots of published neural crest marker genes across each subcluster. **d** Average and percent expression for the top 5 transcription factor (TF) marker genes for each

of the CNCC subclusters. **e** Gene expression spatial feature plot for indicated CNCC marker genes in the more medially localized section (Slice2) from a CS13 human embryo⁷¹. **f** Spatial feature plots for modules scores calculated from the top 100 marker genes from indicated CNCC subtypes in the same section as in this figure (e).

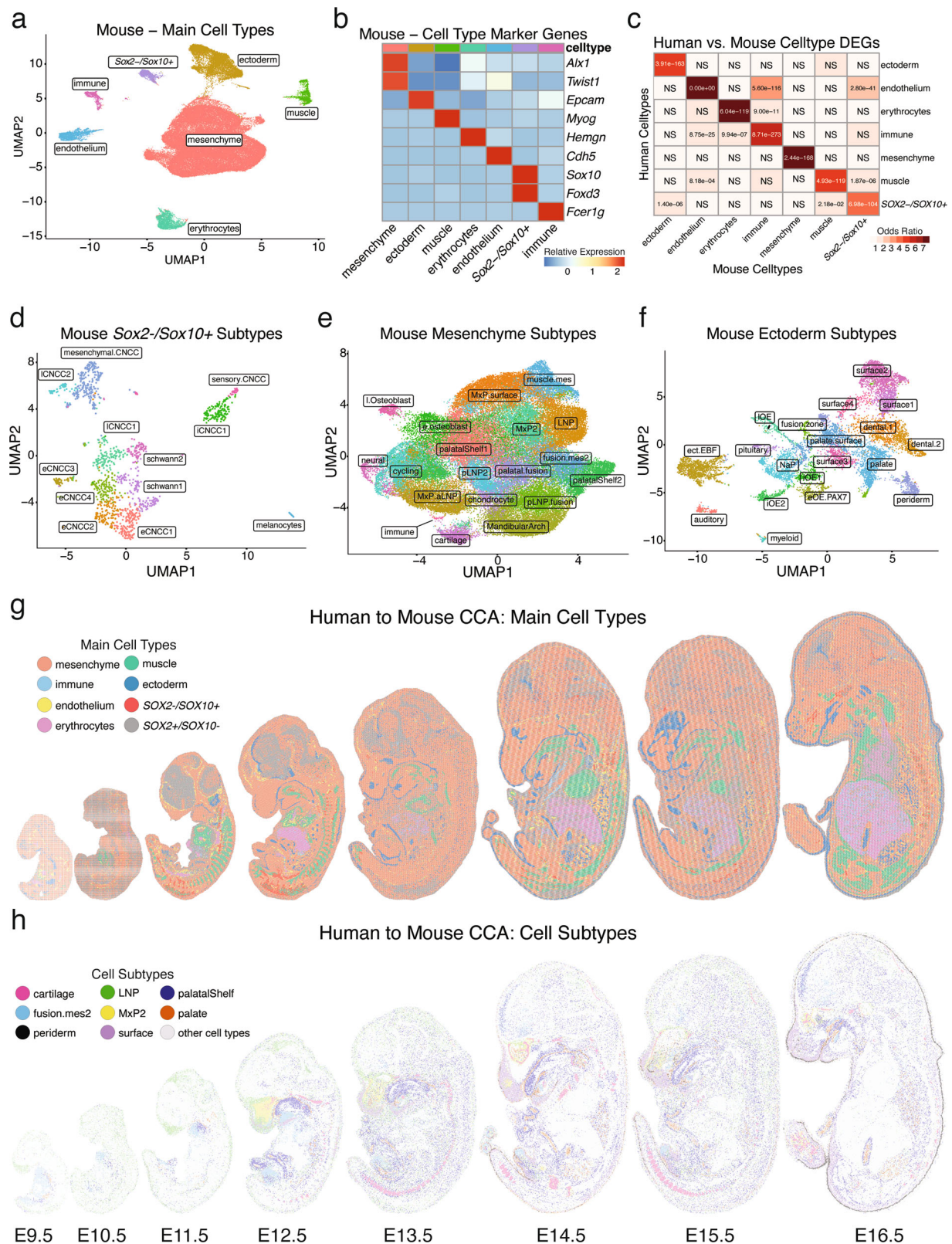


Fig. 4 | Single-nucleus gene expression in the developing mouse face. a UMAP projection of all cells profiled in this study and published studies by main cell type. **b** Heatmap of expression for indicated marker genes across each cluster. **c** Heatmap of sharing of marker genes between each major cell type in human and mouse (p -values calculated by GeneOverlap R package¹³⁸). **d** UMAP projection of

subclustered CNCC cells from mouse. **e** UMAP projection of subclustered mesenchymal cells from mouse. **f** UMAP projection of subclustered ectodermal cells from mouse. **g** Spatial predictions of human craniofacial main cell types projected on mouse E9.5–16.5 embryo sections⁷¹. **h** Spatial predictions of selected human craniofacial subtypes on mouse E9.5–16.5 embryo sections⁷¹.

subtypes (Fig. 4d and Supplementary Fig. 13). When examining the functional enrichment of marker genes of these subclusters, we found similar results as in human, including a clear population of melanocytes (Supplementary Fig. 14 and Supplementary Data 6). Examination of the same neural crest markers as in human CNCC subtypes revealed very similar patterns of expression with *Sox10* and *Nr2f2* expressed across all the subtypes as well as a similar trend in expression of *Foxd3*, *Pax3*, *Tfap2a*, and *Tfap2b* across subtypes (Supplementary Fig. 15A). When we inspected the transcription factor markers for each of these subtypes, we found many similarities to human including *SOX5* in early CNCC's, *NR2F1* and *NR2F2* in intermediate CNCC's, *TWIST2* in mesenchymal CNCC's, and *NEUROD1* and *NEUROG1* in sensory CNCC's (Supplementary Fig. 15B and Supplementary Data 7). We reprocessed mouse E11.5 spatial transcriptomic data⁷¹ ("Methods") and found similar expression patterns for many of the human CNCC markers in mouse tissues (Supplementary Fig. 16A). Examination of module scores calculated from the mouse CNCC subtypes also revealed similar patterns across the mouse embryo as observed for human (Supplementary Fig. 16B). To identify the orthologous CNCC subtypes across species we compared sharing of orthologous marker genes as we did with the main cell types, as well as an orthogonal method, SAMap⁷², which compares cell types with shared expression programs across species. Both approaches showed highest similarity amongst the human and mouse mesenchymal-committed putative CNCC's as well as human and mouse melanocytes (Supplementary Fig. 17A, D).

We then subclustered the large mouse mesenchyme cluster and identified 19 subclusters across the mouse timeseries (Supplementary Fig. 18A). Using a combination of gene ontology enrichments of marker genes and previous annotations of mouse craniofacial single cell and spatial transcriptomics, we assigned functional and/or positional labels to each cluster (Fig. 4e, Supplementary Data 8 and 9). For example, the well-established lateral nasal process (LNP) marker *Pax7*^{73–75} and the osteoblast marker *Sp7*^{76,77} were used to define respective clusters (Supplementary Fig. 18B). We observed two clusters expressing similar osteoblast markers but were biased in cells from different stages of development, thus, we further refined these as early and late osteoblasts (Supplementary Fig. 18B, C). When we examined mouse E11.5 spatial transcriptomics data, we found good concordance between marker gene expression and generalized localization in the embryo (Supplementary Fig. 19A). Furthermore, projection of module scores for mouse mesenchymal clusters readily identified specific regions of the developing craniofacial structures that corresponded well to their assigned labels (Supplementary Fig. 19B).

We performed identical analyses for the ectodermal cluster revealing an additional 19 subclusters (Fig. 4f and Supplementary Fig. 20A). Applying the same analysis of marker genes from the literature, gene ontology enrichments, and expression in mouse single cell transcriptomics data, we annotated each of these clusters with functional and spatial labels (Supplementary Data 10 and 11). We identified highly specific ectodermal populations like periderm marked by *Gabrp*^{78,79} and *Grhl3*⁸⁰, cells that will form structures of the inner ear marked by *Oc90*^{81,82}, palate ectoderm identified by *Foxe1*⁸³, and the putative pituitary progenitors marked with *Lhx3* and *Lhx4*^{84,85} among others (Supplementary Fig. 20B). Several of the ectodermal subtypes showed stage-specific biases (Supplementary Fig. 20C). As with the mesenchyme, we found the spatially resolved expression of marker genes corresponded well to expected positions of the mouse embryo (Supplementary Fig. 21A). Module score calculations for each subtype resulted in refined spatial identification of subtypes that matched the labels and expected positions well (Supplementary Fig. 21B). Mouse mesenchyme and ectoderm subtype labels were further confirmed and validated using multiple databases ("Methods"; Supplementary Fig. 22).

We next sought to predict the spatial localization of main cell types identified in both human and mouse by projecting them onto previously published spatial transcriptomics datasets of the whole mouse embryo from E9.5 to E16.5⁷¹ and the E15.5 mouse head^{23,86}. We found expected patterns of cell type localization for human cell types projected onto the mouse that reflect known anatomical structures (Fig. 4g). Additionally, we projected these cell types on our recent analysis of spatial gene expression in mouse E15.5 craniofacial sections²³ and similarly found expected patterns of cell type localization (Supplementary Fig. 23A, B). Overall, the analyses performed here confirmed the identities of multiple cell types and subtypes across the development of mouse craniofacial tissues. Moreover, the demonstration of conserved marker genes provides a framework for transferring cell type labels to subclusters identified in human data, as well as putative spatial inferences from data that originally lacked that information, which we explore below.

Characterization of human mesenchymal cell subtypes

We then subclustered the human mesenchymal population and identified 25 subtype clusters (Supplementary Fig. 24A). We applied the same methods of assigning cell types and/or functional labels to each of the human clusters using the gene marker approach based on orthologous gene expression in addition to our orthogonal method using SAMap (Fig. 5a and Supplementary Fig. 17B, E). In some cases, multiple human clusters correlated well with a single mouse cluster and were labelled as separate populations (e.g., mouse palatalShelf2 and human palatalShelf2.4). The most abundant cell types were obtained from the LNP and the fusion mesenchyme (fusion.mes2) and were well represented from each of the timepoints. Some of the transient structures like the muscle mesenchyme (muscle.mes) and cells labeled as palatal.fusion were biased towards early timepoints, while later forming cell types and structures such as cartilage and palatal shelves were dominated by cells derived from CS20 samples (Fig. 5b and Supplementary Fig. 24B). When we examined the transcription factor marker genes identifying each of these clusters we found *MSX2* in the fusion mesenchyme 2 cluster (fusion.mes2), *TFAP2B* in the lateral nasal process population (LNP), *SPX* in the palatal fusion zone population (palatal.fusion); *HOX* genes in the posterior lateral nasal process 2 population (pLNP2); *MKX* in palatal shelf population 1 (palatal.shelf1); and *MYF5* in the muscle mesenchyme population (muscle.mes), among many others (Fig. 5c, Supplementary Fig. 24C, and Supplementary Data 12).

Gene ontology analysis revealed subtype-relevant enrichment, such as bone morphogenesis and ossification in osteoblasts, cartilage development in cartilage, and embryonic organ development and regionalization in fusion.mes2 populations (Supplementary Fig. 25 and Supplementary Data 13). Examination of disease enrichments across cluster marker genes revealed some tissue-specific disorders like Osteogenesis imperfecta in late osteoblasts and Chondrodysplasia in cartilage. Enrichment for genes related to isolated cleft palate were found in several clusters, including cartilage, fusion.mes2, LNP, MandibularArch.1, palatalShelf1, palatalShelf2, and palatalShelf2.2 (Supplementary Fig. 25F).

We next examined some of the markers that defined the mesenchyme versus other cell types, such as *TWIST1* and *SATB2* in the CS13 human spatial transcriptomics data. We found fairly broad expression of *TWIST1* across the embryo with some bias toward the craniofacial region while, *SATB2* was much more regionally restricted to the pharyngeal arches (Fig. 5d and Supplementary Fig. 26A). We examined subtype marker genes including *BARX1*, a marker of MxP.aLNP, and *DLX5*, a marker of the MandibularArch, were rather specifically localized in the general region of the pharyngeal arches, while *BARX1* was also expressed in the developing stomach. *EMX2*, a marker of the MandibularArch, and *DMRTA2*, a marker of muscle.mes,

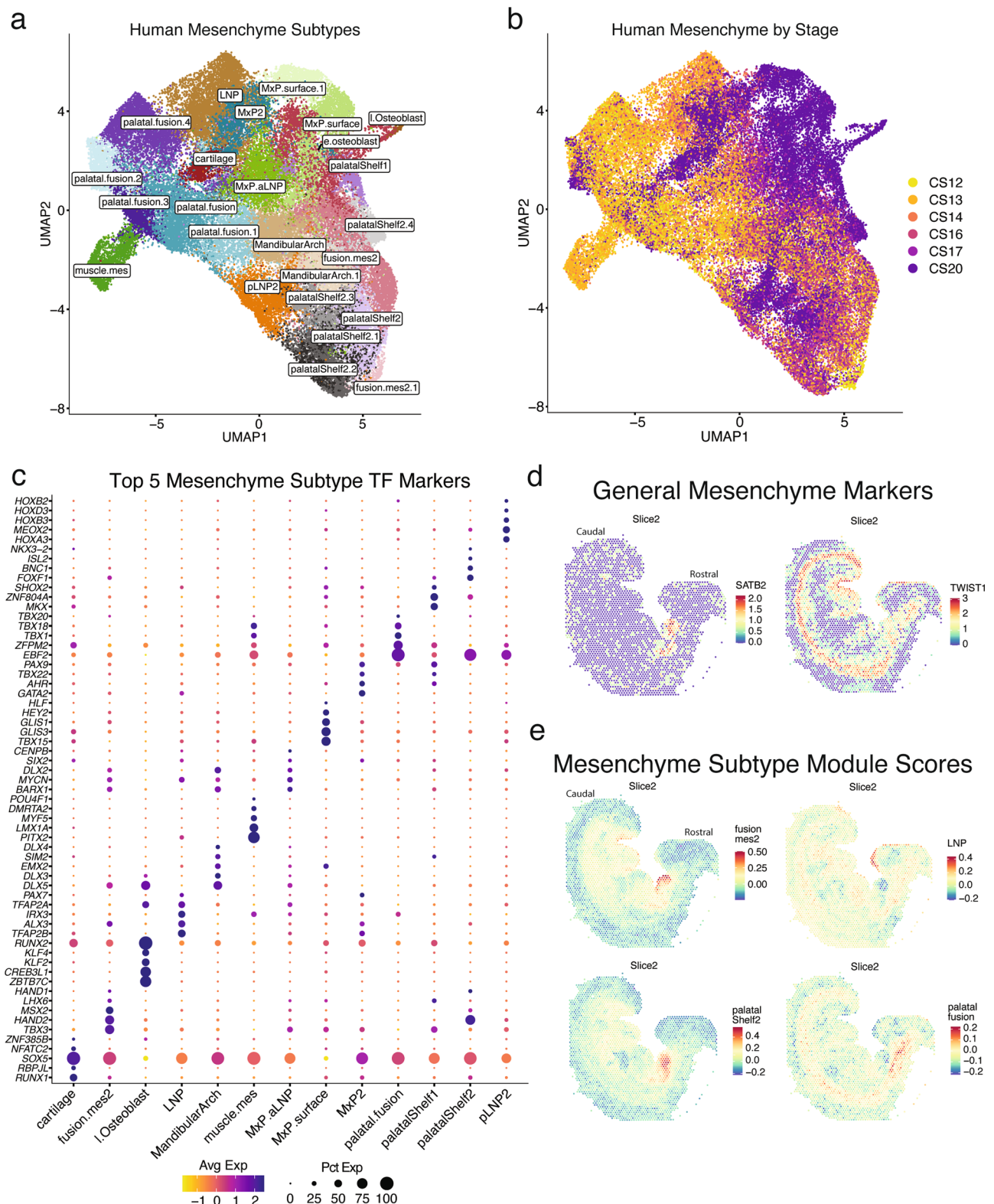


Fig. 5 | Identification of mesenchymal subtypes in human craniofacial development. **a** UMAP projection of subclustered mesenchymal main cell type. Subtype labels based on transfer of mouse mesenchymal subtypes to human. **b** Distribution of cells from each stage across the mesenchymal subclustered UMAP projection. **c** Average and percent expression for the top 5 transcription factor (TF) marker

genes for each of the mesenchymal subclusters. **d** Gene expression spatial feature plot for indicated mesenchymal marker genes in the more medially localized section (Slice2) from a CS13 human embryo⁷¹. **e** Spatial feature plots for modules scores calculated from the top 100 marker genes from indicated mesenchymal subtypes in the same section as this figure (**d**).

were both more restricted to the head region of the embryo at this stage (Supplementary Fig. 26B).

As was observed in mouse, we found much more regionalized signals from module scores for each subtype. The fusion.mes2 cluster were clearly enriched in the pharyngeal arch region of the CS13 embryo and biased toward the more anterior portion of this region (Fig. 5e). The LNP cluster was enriched in distinct areas of the head near the frontonasal region, while the MxP2 and other palatal shelf clusters showed good spatial concordance with the labels that had been assigned (Supplementary Fig. 26C). When we projected these subtypes onto mouse spatial data, we found good concordance for cell subtype labels and known anatomical features (Fig. 4h). Additionally, we projected these subtypes on our recent analysis of spatial gene expression in mouse E15.5 craniofacial sections²³ and similarly found expected patterns of localization (Supplementary Fig. 23C, D).

Characterization of human ectodermal cell subtypes

Using the same approach as described for the mesenchyme, we identified 22 distinct ectodermal clusters (Fig. 6a and Supplementary Fig. 27A). Transferring of mouse labels (Supplementary Fig. 17C, F) revealed cells that give rise to specific ectodermal-derived organs like the pituitary and thyroid, structures of the inner ear (auditory), and surfaces of several structures including periderm (Fig. 6a). As with mesenchymal subclusters, many of the ectodermal subclusters displayed stage-specific bias (Fig. 6b and Supplementary Fig. 27B). Amongst transcription factor marker genes of ectodermal subtypes, *GATA6* was enriched in the palate.1 subtype, *SP8* and *NR2E1* in the nasal placode (NaP), *SALL3*, *TBX2*, and *LMX1A* in the auditory subtypes; *PAX1* in the fusion.zone cluster, among others (Fig. 6c, Supplementary Fig. 27C, and Supplementary Data 14).

Consistent with our findings for CNCCs and mesenchyme, gene ontology analyses revealed subtype relevant enrichment including inner ear morphogenesis and development for auditory subtypes, structural components of the lens in the eye cluster, cornified envelope and skin development in the periderm cluster, thyroid hormone synthesis in the thyroid progenitor cluster, and epithelial to mesenchymal transition in the EMT subcluster (Supplementary Fig. 28 and Supplementary Data 15). Ectodermal surface clusters were enriched for a variety of categories with palate.surface.3 marker genes biased for oxidative phosphorylation and salivary gland development (Supplementary Fig. 28). Examination of enriched human diseases revealed many tissue- or region-specific disorders including microcornea and ectopia lentis in the eye cluster, nonsyndromic deafness in auditory clusters (auditory and auditory.1), thyroid agenesis for the thyroid progenitor cluster, anterior pituitary hypoplasia for the pituitary progenitor cluster (Supplementary Fig. 28F).

Examination of ectodermal markers *GRHL2* and *ESRP1*, revealed relatively restricted expression to various surfaces in the human spatial transcriptomics data, while *OC90* was strongly expressed in putative inner ear (Fig. 6d and Supplementary Fig. 29A). Subtype markers also showed generally restricted expression particularly for *SIX2*, *NR2E1*, *SIX6*, and *NKX2-1* (Supplementary Fig. 29B). When we inspected module scores of each subtype, we observed exquisitely specific localization for some clusters like eye and auditory. Other clusters were generally enriched at surfaces of the pharyngeal arches and the putative esophagus (Supplementary Fig. 29C). As we observed for mesenchyme subclusters, projection onto mouse spatial data revealed known anatomical features (Fig. 4h).

Cell type specific contributions to craniofacial phenotypes

The analyses above demonstrated strong concordance between human and mouse cell types and showed strong support of our cell type labelling, giving us the opportunity to interrogate the cell type-specific expression profiles for enrichment of craniofacial related genetic signals. The genetic contributions of common variants to

many aspects of craniofacial variation have been studied in multiple populations based on frontal and profile photographs^{87,88}. However, the cell types and embryonic landmarks that drive these differences are currently unknown. Following uniform processing of GWAS summary statistics, assigning variants to genes, and calculating expression weighted cell type enrichments ("Methods"), we found that measures likely to be influenced by bone or cartilage structure such as jaw, chin, and brow protrusion as well the positioning of the eyes relative to the base of the left nostril (EnL-AIL) and the positioning of the base of the left nostril to the base of the right nostril (AIL-AIR) were enriched primarily amongst mesenchymal subtypes (Fig. 7a and Supplementary Fig. 30). For profile landmark measures, we found metrics related to soft tissues including multiple measures of lip thickness and shape, ear size, and nose shape were enriched primarily in ectodermal subtypes (Supplementary Fig. 31). Interestingly, while CNCCs give rise to many downstream cell types and tissues, we found relatively few shape associations for putative CNCC subtypes. Overall, this analysis suggests specific cell subtypes contribute differentially to facial variation and suggest these effects begin to manifest early in human development.

Next, we sought to determine if these cell subtypes might be linked to human disease. We posited that integrating continuous expression patterns in an unbiased fashion might yield phenotypic associations reflective of their anatomical location or function. To address this, we employed a systematic examination of the entire Human Phenotype Ontology (HPO)⁸⁹⁻⁹¹ (Fig. 7b). The ectoderm cluster was enriched for abnormality of the integument (158 of 686 phenotypes) and abnormality of head or neck (98 of 1056 phenotypes) (Supplementary Fig. 32). Amongst subtypes we found the eye.ect subcluster was strongly enriched for phenotypes related to abnormalities of the eye (78 of 731), the periderm was most enriched for abnormalities of the integument, and the ect.GDNF subtype was significantly associated with abnormalities of the voice. Among the mesenchymal subclusters, many were enriched for abnormalities of the head and neck as well as abnormalities of the musculoskeletal system. The cartilage cluster showed the most diverse enrichments, including phenotypes related to growth abnormalities and abnormalities of the musculoskeletal system and limbs. The *SOX10*⁺ putative CNCC cluster was most enriched for abnormalities of the nervous system, while the *SOX2*⁺ early progenitor cluster was most enriched for abnormalities of the eye. Together, these results suggest that some subtypes show phenotypic specificity, while others are more generic.

We next turned to studies of the genetic underpinnings of craniofacial abnormalities, particularly, the GWAS that have been systematically performed on a large cohort of Finnish ancestry⁹² to determine the cell types that potentially influence risk for orofacial clefting. When we analyzed these GWAS using the same approach as for facial variation we found similar partitioning of enrichment between specific classes of cell subtypes (Fig. 7c). The immune cluster was most significantly enriched for Crohn's and Systemic Lupus Erythematosus, serving as a good control for the specificity of this analysis (Supplementary Fig. 33). We found that cell subtypes were enriched for ankyloglossia, other congenital malformation of the tongue and mouth, and cleft palate (Fig. 7c and Supplementary Fig. 33). Specifically, we identified ectoderm subtype annotated as palate.3 as the most significantly enriched for associations with cleft palate. For mesenchymal subtypes we found significant associations with Cleft Lip or Lip and Palate in mesenchymal subtypes related to general fusion, palatal fusion, and the maxillary process surface (Fig. 7c). A few of these subtypes, particularly the palatal fusion subtypes, were associated with both craniofacial disease and normal facial variation suggesting their dual contribution to facial shape as well as risk for clefting. When we examined the spatial mapping of the risk associations of Cleft Palate and Cleft Lip and Cleft Palate using gsMap⁹³

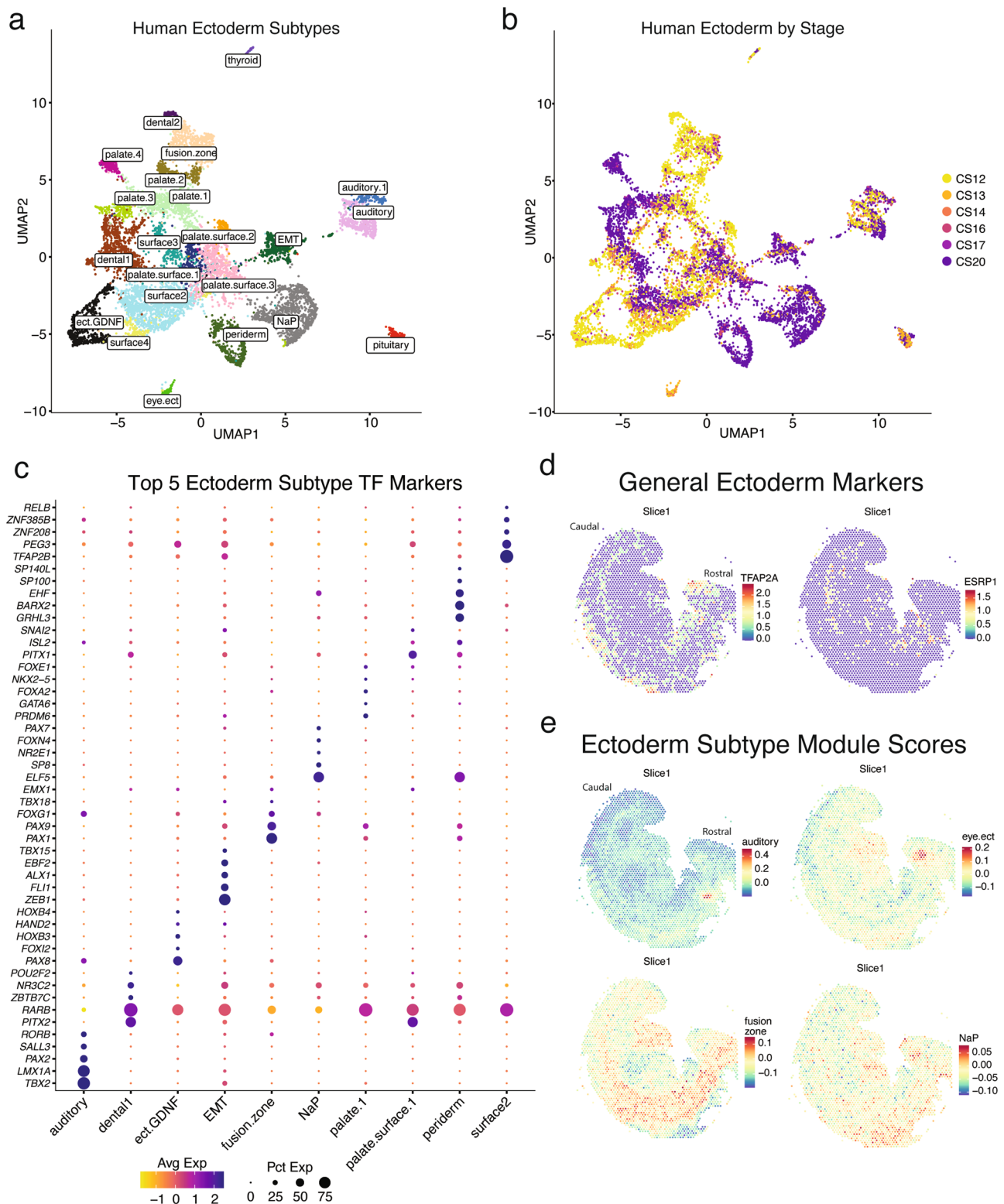
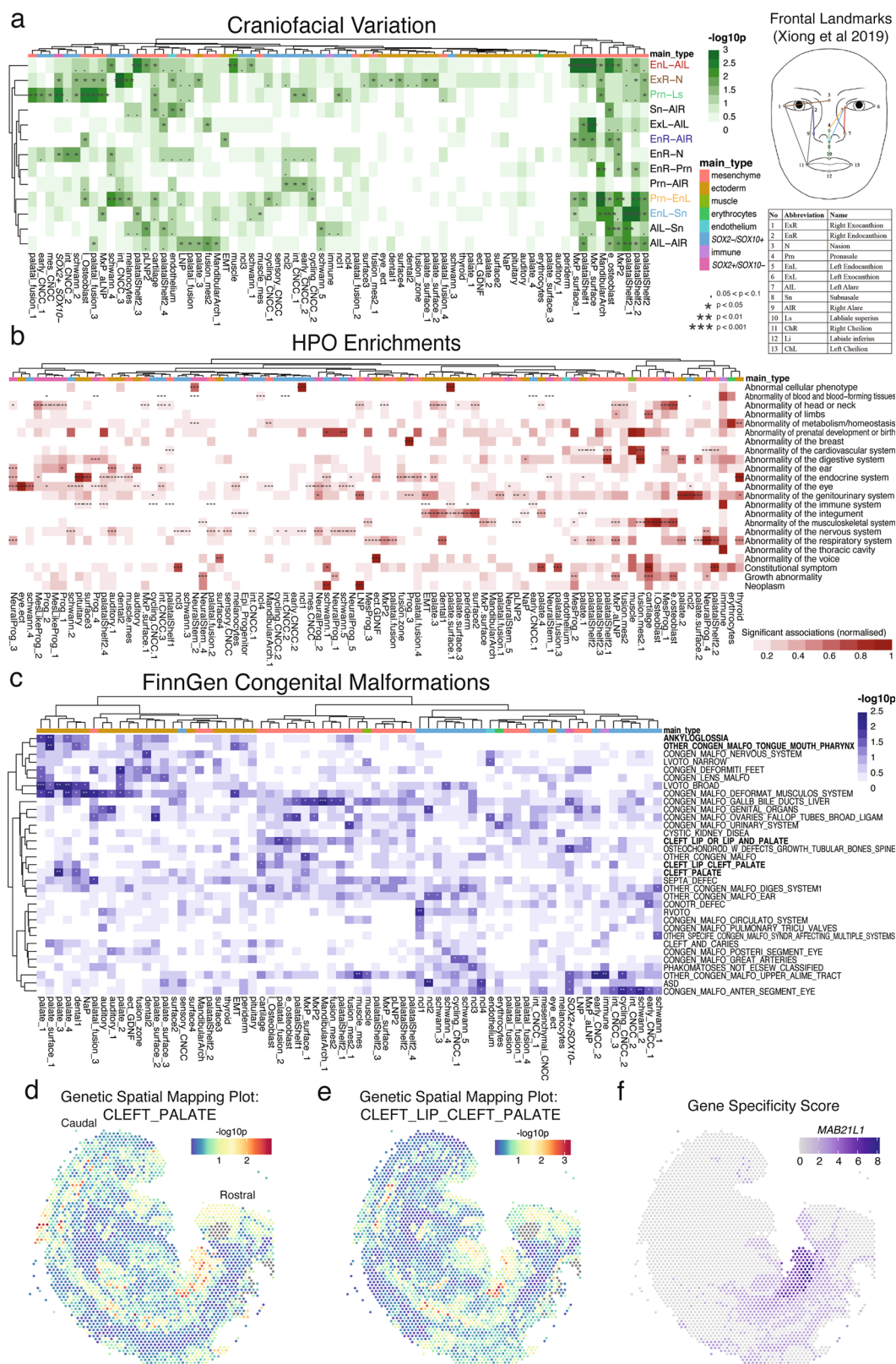


Fig. 6 | Identification of ectodermal subtypes in human craniofacial development. **a** UMAP projection of subclustered ectodermal main cell type. Subtype labels based on transfer of mouse ectodermal subtypes to human. **b** Distribution of cells from each stage across the ectodermal subclustered UMAP projection. **c** Average and percent expression for the top 5 transcription factor (TF) marker

genes for each of the subclusters. **d** Gene expression spatial feature plot for indicated ectodermal marker genes in the more laterally localized section (Slice1) from a CS13 human embryo⁷¹. **e** Spatial feature plots for modules scores calculated from top 100 marker genes from indicated ectodermal subtypes in the same section as this figure (**d**).

(“Methods”), we found significant associations in the pharyngeal regions of the CS13 human embryo Slice 1 section (Fig. 7d, e) agreeing well with the localization of the subtype module scores (Fig. 5e, and Supplementary Fig. 29C). One of the strongest correlations with

specificity of expression driving the spatial association of the Cleft Lip and Cleft Palate was *MAB21L1* (Fig. 7f), a gene we previously implicated as a putative craniofacial disease-associated gene¹³. As a control, we identified localization of the narrow left ventricular outflow tract



obstruction (LVOTO Narrow) trait and highly correlated *B4GALNT3* gene specificity to the putative heart region (Supplementary Fig. 34).

Cell type specific enrichment of disease relevant gene lists

Thus far our analyses of the craniofacial cell types have only leveraged annotated ontology categories and common variant associations, but do not include other gene lists that may be of use to the craniofacial

field. To address this, we assembled multiple gene lists relevant for orofacial clefting including those compiled by CleftGeneDB⁹⁴, genes co-expressed in important gene modules or prioritized for craniofacial disease in our recent work¹³³, and genes with distinct classes of de novo mutations (synonymous vs protein-altering) in orofacial cleft trios sequenced as part of the Gabriella Miller Kids First program^{95,96} and CPSeq Studies⁹⁷. We also curated genes at the extremes of

Fig. 7 | Enrichment of common variation associate with facial shape differences and congenital abnormality risk. **a** Clustered heatmap of significance values calculated by MAGMA.Celltyping¹⁵² for select significant facial variation traits and cell subtypes. Frontal landmark diagram adapted from Xiong et al.⁸⁷. Colors along top of heatmap indicate main cell type classification. Trait measures are obtained from Xiong et al. 2019 and combinatorial code is indicated in coded legend (e.g., EnL-AIL indicates landmark segment 5 to 7). Levels of significance indicated by asterisks or period according to figure. **b** Heatmap showing number of enriched human phenotypes (max-normalized from 0 to 1 within each branch) for cell subtypes as calculated by MSTExplorer::run_phenomix. Significance of the proportion

tests, testing for disproportionate numbers of phenotype enrichments for a given cell subtype within a given HPO branch, is denoted with asterisks (FDR < 0.001 = ***, FDR < 0.01 = **, FDR < 0.05 = *). **c** Clustered heatmap of significance values calculated by MAGMA.Celltyping¹⁵² for select significant FinnGen⁹² congenital abnormalities or diseases and cell subtypes. Colors along top of heatmap indicate main cell type classification. Levels of significance indicated by asterisks or period according to figure legend. **d** Spatial mapping of GWAS variants associated with Cleft Palate on CS13 spatial transcriptomics as calculated by gsMAP⁹³. **e** Same as this figure (**d**) but for Cleft Lip and Cleft Palate GWAS. **f** Gene Specificity Score for *MAB21L1* on CS13 spatial transcriptomics.

tolerance to loss of function mutations in otherwise healthy populations that have been suggested to be enriched or depleted of disease-relevant genes⁹⁸. Lastly, given our findings for common facial variation across humans, we wondered whether genes potentially regulated by Neanderthal derived sequences might have craniofacial cell type specific enrichments. As a control for this evolutionary analysis, we included genes near human accelerated regions (HARs), which have been reported to be enriched in neuronal related functions and expression^{99–102}.

We again employed the EWCE approach and found that the CleftGeneDB, craniofacial black co-expression modules, and our prioritized gene lists showed similar patterns of significant enrichments in mesenchymal subtypes, including multiple clusters related to the maxillary process, palatal shelves, and LNP (Fig. 8a). Additionally, the genes identified by gnomAD to have the least tolerance for loss of function mutations (LOUEF decile 1) were also significantly enriched in these mesenchymal subtypes. This contrasted with those genes identified by gnomAD with the most tolerance for loss of function mutations (LOUEF decile 9) that showed few enrichments and were generally non-overlapping with the LOUEF decile 1 enrichments. Interestingly, relatively few ectodermal and CNCC subtypes were enriched for these gene lists.

When we analyzed the genes near Neanderthal-derived sequences, we found patterns of cell type enrichment distinct from the more disease-focused lists described above. The specialized ectodermal EMT and ect.GDNF clusters as well as both mesenchymal maxillary process surface clusters (MxP.surface and MxP.surface.1), multiple palatal shelf clusters, and both auditory subtypes were enriched in this analysis. This was contrasted by only a single progenitor cell subtype enriched in the HAR-associated gene list, consistent with their previously published association with brain cell types and neuronal function^{99–101}. We found no significant enrichments from any of our randomly selected gene lists (“Methods”) across cell subtypes and only found enrichment for the gnomAD decile 9 genes in erythrocytes and immune cell types (Fig. 8a). We then turned to recently identified de novo variants from orofacial clefting trios from the Gabriella Miller Kids First program^{95,96} and CPSeq studies⁹⁷. We found no enrichment in any cell types for genes affected by de novo synonymous variants (Fig. 8a). In contrast, we identified multiple ectodermal cell subtypes that strongly express genes with de novo protein altering variants, including palate.1 and NaP, as well as multiple palatal shelf mesenchymal subtypes.

We finally explored whether the total number of de novo variants observed in genes might reveal additional disease associations. We applied a computational framework that examines gene lists for excess de novo mutational load^{103,104} and identified 29 clusters for which at least one phenotype was significantly enriched for protein-altering variants using a Benjamin-Hochberg false discovery rate of <10%¹⁰⁵ (Fig. 8b and Supplementary Data 16). These include 27 subtypes for all trios with orofacial clefts (OFCs), 21 subtypes for trios with CLP, and 7 subtypes for trios with CP. We identified 16 significant enrichments across ectodermal cell types ($n = 22$), 5 enrichments from mesenchymal cell types ($n = 25$), and 8 enrichments from the early progenitor

population ($n = 19$). Only 3 clusters were significantly enriched in all three categories (palate.surface.3, EMT, and NeuralProg.5), none of which were mesenchymal subtypes, whereas there were 16 shared between all OFCs and CL/P and 4 shared between all OFCs and CP. We also found 2 clusters that were only significant in the CL/P group (pituitary and MxP.surface.1) and 4 clusters that were only significant in the full cohort (NeuralProg.4, MesenchymalLikeProg.1, NeuralProg.2, auditory). We identified strongest enrichment for de novo variants identified in the whole cohort and those probands with cleft lip and/or cleft palate (CL/P) in the ectodermal subtypes: palate.surface.3, EMT, palatal.surface.1, NaP, palate.3, surface3, among others. Interestingly, several subtypes were biased toward significant enrichment related to CP vs CL/P. For instance, palate.1, Prog.1, Prog.2, and cartilage were enriched for the former, while both dental clusters, and multiple palate and palate surface-related clusters were enriched for the latter (Fig. 8b).

For de novo variants identified in probands with CL/P from significantly enriched ectodermal subtypes ($n = 14$), the main driver genes were *ESRP1*, *GRHL1*, *GRHL2*, and *TPD52*. Variants identified in probands with CP from significantly enriched ectodermal subtypes ($n = 3$), additionally included *CDH1*, and *LLGL2*. These genes have fairly restricted expression in the head region and presumptive somites of the CS13 embryo (Supplementary Fig. 35). This enrichment highlights the importance of these cells in OFC etiology, but the difference in signal drivers may provide insight into the heterogeneity of the genetic architecture between CP and CL/P.

Discussion

Craniofacial abnormalities are some of the most common human birth defects. Only recently have gene expression patterns active during human craniofacial development been examined¹³. We previously showed that genes specifically or co-expressed across craniofacial development relative to other tissues were enriched for known disease-causing genes¹³. However, these analyses relied on bulk gene expression data from the developing craniofacial tissues. Our bulk analyses showed strong bias for gene programs expressed in human and mouse mesenchyme preventing analysis of genes in less abundant cell types. While other single cell atlases from human embryonic development have been described, there were few biological replicates and relatively few cells clearly derived from craniofacial regions^{24–27}.

By generating comparable datasets from both mouse and human, we had the unique opportunity to identify both shared and species-biased gene programs active in individual cell types. As expected, we found the main cell types identified in each species share the most significant amount of marker genes with the orthologous cell type in the other species. Among these, the endothelium, mesenchyme, and erythrocyte clusters were the most functionally shared between human and mouse, with the CNCC cluster exhibiting the least conservation. The notable exclusion was the *SOX2*+/ *SOX10*– putative progenitor population. These were not identified in the mouse analysis either due to them being much less abundant in mouse tissues, or they are derived from extraneous tissues like the developing brain from the

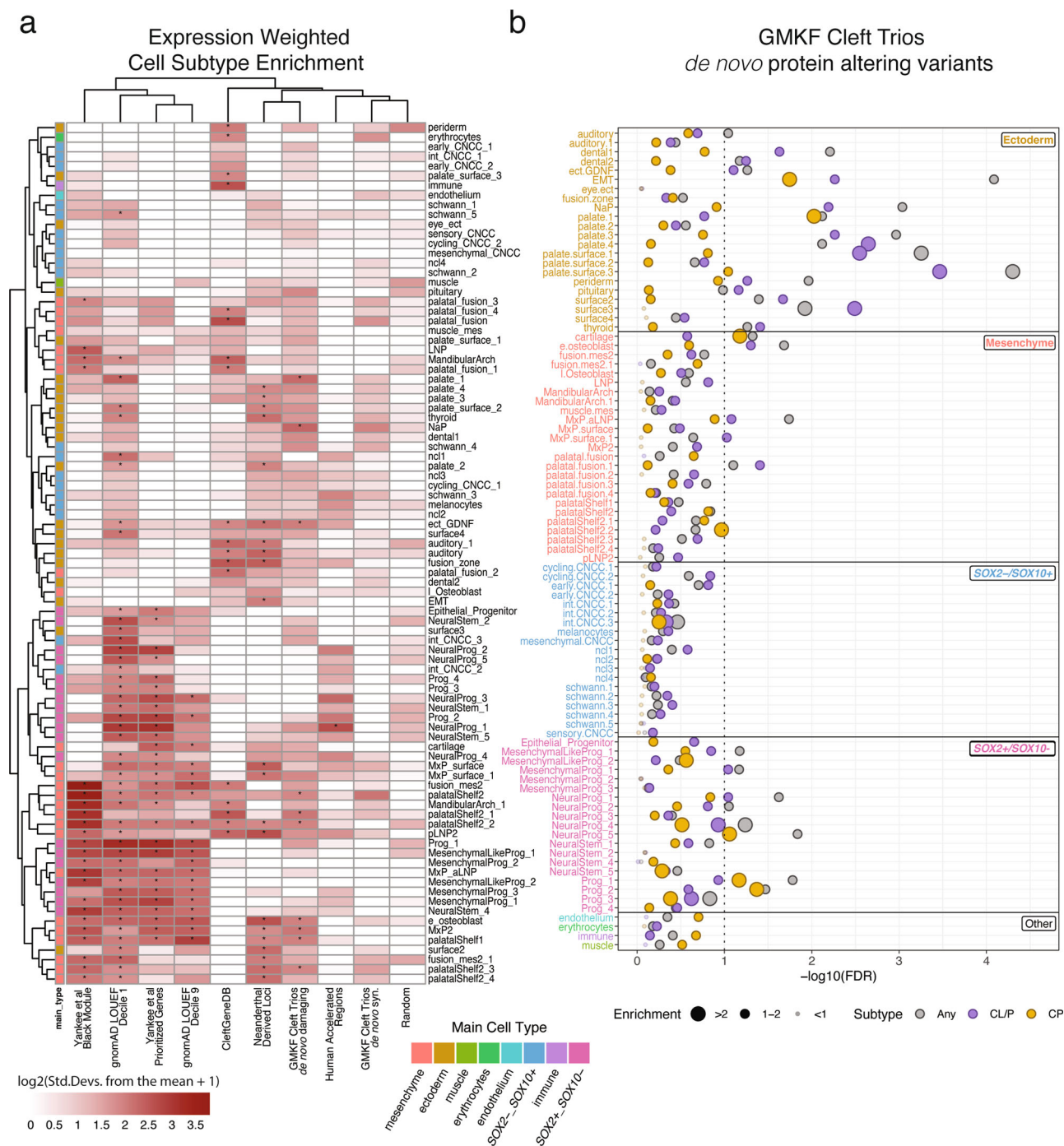


Fig. 8 | Genes associated with orofacial clefting, constraint in human populations, and Neanderthal introgression show distinct cell subtype enrichments. **a** Heatmap of standard deviations from the mean of bootstrapping tests performed by EWCE method⁵¹ for each indicated gene list and cell subtype. Asterisks indicate significant subtype enrichments corrected for number of gene lists and cell subtypes performed. **b** Bubbleplot of $-\log_{10}$ transformed significance and fold

enrichment values for each cell subtype from denovoZer¹⁰³ analysis of protein *damaging* *de novo* variation in orofacial cleft trios from the Gabriella Miller Kids First program^{95,96}. Colored circles indicate variants identified in whole cohort (Any), in cleft lip with cleft palate probands (CL/P), or cleft palate only probands (CP). Cell subtypes are clustered by main cell type.

human samples and could not be removed by any of our ad hoc filtering steps. The decreased conservation of CNCC gene expression programs could indicate this cell type may be particularly labile across evolution allowing innovation of craniofacial shape as others have proposed^{118–20,66,106}. Alternatively, these differences could be reflective of our narrow definition of CNCCs or technical differences in the non-CNCC subclusters obtained in each species. Further analysis of differentially expressed genes from these subtype clusters in developing

mouse embryos and human culture models could help to disentangle these possibilities.

In this work, we also leveraged the substantial single-cell and spatial transcriptomics resources and cell type annotations for mouse that have been generated by many different groups^{23,107}. Using these data sources, we were able to systematically transfer functional and spatial labels for mouse cell subtypes to our human data. We confirmed these labels using gene marker and disease ontology analyses in addition to

leveraging previously published spatial transcriptomics data for a CS13 human embryo²⁵ and across multiple sections of developing mouse embryos. These analyses largely confirm anatomical and regional annotations for each subtype and give credence to our human genetic enrichments we discuss below. It will be necessary to further characterize the markers we identified with orthogonal approaches, such as protein expression or subtype labelling experiments to definitively determine the identity of the clusters we have described.

A major goal of generating this atlas was to enable better understanding of how facial shape is encoded in our genomes. In recent years, coupling of two- and three-dimensional imaging approaches with large-scale genotyping has enabled the discovery of common genetic variants associated with quantitative differences in many different facial landmarks^{87,88,108–110}. While these variants have been shown to be enriched in regulatory regions active in the developing face, the cell types that underly facial differences were unknown. Using our cell subtype annotations, we found distinct differences in enrichments for measurements across the human face. In general, the enrichments we observed were mutually exclusive, features likely driven by mesenchyme subtypes were not associated with an ectodermal subtype and vice versa. As expected, mesenchyme subtypes were associated with features that are likely driven by hard structures like bone and cartilage, while ectoderm subtypes were associated with measures related to soft tissue shape or thickness. The most consistent enrichments were related to variation in measures of the midface for mesenchyme subtypes, including the maxillary process, palatal shelves, and fusion zone. While the two studies we utilized were performed in populations with distinct ancestries, it is possible that other subtypes could influence facial variation in other populations. Further identification of genetic associations with more facial measures in a more diverse set of individuals will be needed to address this issue.

The most common form of craniofacial abnormalities, non-syndromic cleft lip and/or cleft palate, is thought to occur between 4 and 6 weeks in human development^{8,111,112}. We found that variants associated with risk for orofacial clefting are enriched in regulatory sequences active in craniofacial tissues from this developmental window²⁸. However, the cell types in which these variants manifest their effects were unknown. Indeed, we found the genes that specify some subtypes of mesenchyme, ectoderm, and early progenitor populations to be significantly enriched for nearby variants associated orofacial clefting or other abnormalities of mouth. Interestingly, several of the enrichments we observed for subtypes were shared across the facial variation and craniofacial abnormality analyses, including cartilage, MxP.aLNP, palate.3, and palatal.fusion.2. This is particularly interesting as it has been speculated that some of the same processes that drive normal facial variation can result in disease^{113–117}. Our analysis of marker genes for the MxP.aLNP subtype revealed plausible localization near the fusion zone termed the “lambdoid junction”^{118–120}. Failure of this region to fuse in humans has been suggested to cause cleft lip that could also involve the nostril region and primary palate^{10,111,121,122}. Thus, subtle differences in the timing of migrations and fusion of cells residing in this region could influence the shape of the midface. Interestingly, some of the major markers of the specialized EMT subtype are the EBF family of transcription factors, which have been linked to regulation of differentiation and predisposition for several tumor types^{123–129}. Our previous work suggested that these genes were co-expressed more strongly in human craniofacial cell types than mouse and found compelling evidence that *EBF3* is a bona fide orofacial clefting risk gene¹³. The timing of differentiation of cells at a fusion zone could influence the degree to which structures fuse and impact both clefting risk and facial shape. Studies leveraging the marker genes we have identified for each of these subtypes could allow for future experiments studying the impact of facial variation and disease risk.

Our analysis of curated gene lists not included in standard gene ontologies revealed cell subtypes likely enriched in craniofacial

diseases. For instance, our previous prioritized gene list¹³ as well as the curated CleftGeneDB resource are heavily biased toward mesenchymal subtypes, including the maxillary process, palatal shelves, and LNP. While both the common variant analyses for orofacial clefting and the curated craniofacial disease gene lists were biased toward mesenchyme subtypes, genes harboring rare de novo protein damaging variants identified in cleft probands showed much greater enrichment in ectodermal subtypes. This trend was further supported when we examined the frequency of de novo protein altering variants, where we found significant enrichment for CL/P in multiple ectodermal subtypes. While the number of CP only cases were fewer than CL/P, we found these de novo variants were also enriched in a few ectodermal subtypes (palate.1, EMT, and palate.surface.3), early progenitor subtypes (Prog.1, Prog.2, and NeuralProg.5), and a single mesenchymal subtype (cartilage). Overall, this points to the significance of ectodermal subtypes, that make up a small proportion of craniofacial tissue but play a major role in clefting risk.

While we have endeavored to create a high-quality resource and analyze the resulting data in an unbiased fashion, there are several important caveats and limitations that should be acknowledged. First, our two earliest timepoints, CS12 and CS13, are made up of only female or male samples, respectively. While we did not observe any sex-related bias in any of our main cell type clusters, some subtypes were made up of only one sex. Thus, some of the differences we observed in overall conservation of main cell types, particularly for CNCs, could be influenced by this bias. However, most of the subtypes had a clear counterpart in mouse subclustering that were equally mixed for male and female samples. Second, the nature of the human tissue we received precluded precise dissection, as we can perform in mouse guided by anatomical landmarks of the entire embryo. We attempted to identify and remove extraneous cell types from the human samples using module score-based approaches. However, these could have resulted in over- or under-filtering of cell types. Indeed, we were able to identify subclusters that clearly reflected thyroid and pituitary gland progenitors. However, inclusion of these subclusters as well as immune and erythrocyte clusters, served as excellent negative controls demonstrating the specificity of our analyses. Lastly, the gene expression profiles we generated are based on RNA that is either newly transcribed or retained in the nucleus. The levels of mature mRNA could differ substantially in individual cell types, across the cell cycle, and over developmental time. Further direct comparisons of single-nucleus and single-cell gene expression in the same biological contexts are required to understand these differences.

Despite these caveats, we have demonstrated the biological and disease relevance of the cell types and subtypes identified. In total, this is a substantial resource for understanding the cell types and gene expression patterns that build the human and mouse face. Our analyses revealed relationships between specific cell subtypes and many aspects of human biology, including facial shape and orofacial clefting risk. We also illuminated the contribution of ectodermal subtypes to de novo protein altering variants in orofacial clefting. Finally, to make this resource more widely useful and accessible, we distribute the results via a user-friendly web interface (https://cotneyshiny.research.chop.edu/shiny-apps/craniofacial_all_snRNA_v2/) and deposited in the CellXGene Discover database (<https://cellxgene.cziscience.com/collections/690fb28b-e364-4d3d-8934-bae50ae93857>). Future integration with cell-type-specific chromatin accessibility could reveal specific variants and regulatory regions that encode such phenotypic differences, risk factors, and species-specific biology.

Methods

Human tissue samples

The use of human embryonic tissue was reviewed and approved by the Human Subjects Protection Program at UConn Health (UCHC 710-2-13-14-03) and Children’s Hospital of Philadelphia (IRB 24-022258). Human

embryonic craniofacial tissues were collected via the Joint MRC/Wellcome Trust Human Developmental Biology Resource (HDBR) under-informed ethical consent with Research Tissue Bank ethical approval (18/LO/0822 and 18/NE/0290, project 200225). Donors are asked to give explicit written consent for the fetal material to be collected, and only after they have been counseled about the termination of their pregnancy. Donors do not get anything in return for their donation and receive the same hospital care as those who decide not to donate. Further documentation of all policies and ethical approvals for HDBR sample collection can be found at <https://www.hdbb.org/ethical-approvals>. Tissues were flash-frozen upon collection and stored at -80°C . Upon thawing, the samples were quickly inspected for intactness of the general craniofacial prominences and the whole sample was processed for single nucleus multiomics. As these are not entire embryos, we lack many of the additional anatomical landmarks that are available in intact mouse embryos. Thus, we expect that additional tissues and/or cell types will be present in human samples that can only be removed through computational filtering.

Mouse embryonic tissue samples

The use of mouse embryonic tissues was reviewed and approved by the UConn Health Institutional Animal Care and Use Committee (Protocol AP-2000061-0723). Eight-week-old wild-type male and female C57BL6/J mice were obtained from Jackson Laboratory. Mice were housed according to recommendations by Jackson Laboratory with 12 h light:dark cycle beginning at 7 a.m. The ambient temperature was maintained between 20 and 22 $^{\circ}\text{C}$ and humidity was maintained at 40–60%. Mice were given ad libitum access to food and water. Timed matings were established by the identification of vaginal plugs the morning following the housing of a single male with multiple female mice. Pregnant mothers were euthanized by carbon dioxide narcosis at mid-day either 10, 11, or 12 days after identification of the vaginal plug and embryos were harvested. The staging was confirmed by counting somites and comparing overall morphology to the Theiler Staging Criteria¹³⁰. All embryos from a given litter were combined for individual biological replicates, and at least three biological replicates were collected and processed for each stage. Craniofacial prominences were collected in a very similar fashion to human samples, flash frozen, and stored at -80°C until prepared for single nucleus multiomics.

Single-nucleus multiomics

Primary human craniofacial prominences, excluding the eyes and more anterior tissue from CS12, CS13, CS14, CS16, CS17, and CS20, each stage represented by a minimum of 3 replicates, were obtained from HDBR. Tissue from each embryo was mechanically broken into single-cell suspensions and cells were checked for viability and counted using Trypan blue staining following the 10 $\times$ Genomics protocol for single-cell multiome sequencing using the ChromiumX controller. Samples were sequenced on multiple Illumina NovaSeq runs according to 10 $\times$ Genomics recommendations. Raw fastqs were processed using Cell Ranger ARC¹³¹ (v2.0.2) using hg38 genome and gene annotations provided by 10 $\times$ Genomics. For purpose of this paper, we have focused on gene expression to identify cell types in depth and understand their biological relevance. The chromatin accessibility data was of generally high quality based on reports from Cell Ranger ARC. This data forms the foundation of a separate manuscript that will leverage the cell-type and subtype labels generated and validated in this study.

Primary mouse craniofacial prominences, excluding the eyes and more anterior tissue from E10.5–E12.5 from multiple (3–5 depending on stage) mixed sex C57BL/6J Mus Musculus embryos (Jackson Laboratories) were pooled. Animals were raised and sacrificed in compliance with UConn Health IACUC approval (protocol AP-200061-0723). Samples were mechanically broken into single-cell suspensions, processed for multiome using the ChromiumX controller, and sequenced in the same fashion as for human samples above. Raw fastqs were

processed using Cell Ranger ARC (v2.0.2) using mm10 genome and gene annotations provided by 10 $\times$ Genomics.

Processing of snRNA and identification of major cell types

Raw barcode matrices from each human samples generated by Cell Ranger ARC were utilized for downstream processing to maximize total nuclei. Raw barcode matrices for each sample were individually loaded with Read10X_h5 command in Seurat¹³² (v. 5.2.1). Subsequently, the emptyDrops function within the DropletUtils R package (v. 1.24.0) was used to identify nuclei containing ambient RNA for each sample (emptyDrops FDR < 0.001). Individual sample Seurat objects were then merged into a single object. Percentage of mitochondrial reads were calculated for each cell and filtering was performed to only retain cells with less than ten percent mitochondrial derived. Further filtering was performed based on number of counts per cell ($500 < x < 25000$) and number of genes detected per cell ($500 < x < 7000$), resulting in a total of 130,427 nuclei. Filtered data were normalized with default values and cell cycle scores were calculated using Seurat. Data was scaled based on S and G2M score regression and dimensionality reduction with principal component analysis (PCA) were performed using respective commands in Seurat. The top 2000 variable features were identified, and sample-level batch correction was performed using the harmony R package¹³³ (v. 1.2.3). Nearest neighbors based on harmony corrected embeddings were calculated with up to 30 dimensions and clusters were identified with multiple resolutions from 0.1 to 1 in Seurat. Resolution parameters were selected for clustering based on generated elbow plots and identification of the elbow point to determine the optimal number of clusters. We then performed Uniform Manifold Approximation and Projection (UMAP)³⁶ dimensionality reduction using harmony corrected embeddings in Seurat (dimensions = 30, minimum distance = 0.3). Resulting clusters were inspected for expression of multiple craniofacial markers from Li et al.¹² and marker genes were identified for each cluster. Module scores were generated using the AddModuleScore function in Seurat for neuronal markers and radial glia markers from Eze et al.¹³⁴. Cells from clusters identified with high neuronal or radial glia module were removed (See Supplementary Note), resulting in 95,892 nuclei. The process of normalization, harmonization, and clustering was repeated with remaining cells from all samples. We observed that early samples had consistently higher mitochondrial reads (Supplementary Fig. 2D), however, this was not due to increased apoptosis as tested by generating an apoptosis module score using gene markers from the Hallmark Apoptosis gene set within the human MSigDB Core Collection (Supplementary Fig. 2D). The sex of each embryo was also identified using known sex determination genes for males (*UTY*, *DDX3Y*, *KDM5D*, and *NLGN4Y*) and females (*XIST* and *TSIX*) (Supplementary Fig. 2E), and each of the main cell type clusters contained roughly the same number of male and female samples (Supplementary Fig. 2F). Marker genes for major cell types were identified using FindAllMarkers (logfc.threshold = 0.25, min.pct = 0.1, test.use = “wilcox”, min.cells.feature = 3, min.cells.group = 3, pseudocount.use = 1, and return.thresh = 0.01). The top 100 marker genes for each cluster ($p < 0.05$, ranked by log2fold change versus all other clusters) were then analyzed for gene and disease ontology enrichments using compareCluster in clusterProfiler R package (v. 4.12.6). Cranial neural crest genes were compiled based on markers identified by regulatory network construction in human cultured CNCC and craniofacial tissue data⁶⁵. Major cell type labels were applied to each cluster.

RNA velocity analyses. RNA velocity analyses were performed on the main cell types to assess their transcriptional dynamics. We used the “run” command within the velocity python package¹³⁵ (v. 0.17.17) with the sorted bam files from each sample and the human Gencode v32 gtf file as input. This resulted in the computation of a loom file for each sample containing two count matrices: one containing

unspliced counts and another containing spliced counts. The loom files for each of the samples were merged and subset to exclude the previously filtered out cells. We then used the scVelo package (v. 0.3.3) in Python to further filter, normalize, and process the data. We used the `filter_and_normalize` function with parameters `min_shared_counts` set to 20 and `n_top_genes` set to 2000. Next, we computed the moments using `moments` function with `n_pcs` and `n_neighbors` set to 30. Then, the `recover_dynamics` function was run with `n_jobs` set to 30. Using the CellRank package¹³⁶ (v. 2.0.6) in python, we set up the VelocityKernel containing the scVelo-computed velocities to compute a transition matrix using the `compute_transition_matrix` function. To increase the robustness of this calculation, we combined the VelocityKernel with the similarity-based ConnectivityKernel. To accomplish this, we also used the `compute_transition_matrix` function with the ConnectivityKernel and then combined the two kernels using the following function: `combined_kernel = 0.8 × VelocityKernel + 0.2 × ConnectivityKernel`. We visualized the transition matrix on the UMAP of all main cell types using the `plot_projection` function.

For mouse data sets, filtered barcode matrices from each mouse E10.5 to E12.5 samples generated by Cell Ranger ARC were individually loaded with `Read10X_h5` command in Seurat¹³² and merged with E13.5–E15.5 data from Pina et al.²³ (GSE205448). Calculation of percent mitochondrial reads and filtering were performed with similar thresholds to human data above. Subsequent harmonization, dimensionality reduction, and clustering were performed identically to those for human data above. Less significant contamination of neuronal cell types was observed in mouse data, which was identified and filtered as described for human. Identification of marker genes and gene ontology enrichments, and CNCC modules scores were performed as above for the human data. Major cell type labels were applied to each mouse cluster.

Marker gene comparisons across species

Lists of all marker genes for each of the eight main subtypes for each species ($p < 0.05$) were compiled and orthology based on HGNC symbol = was obtained using the `convert_orthologs` command in the orthogene R package¹³⁷ (v. 1.10.0) with `non121_strategy` set to “drop-both-species” and `method` set to “gprofiler”. Only genes that had one ortholog in each species and a HGNC symbol were retained ($n = 15,699$). Significant overlaps between all orthologous marker gene lists were determined using the `newGOM` command in GeneOverlap R package¹³⁸ (v. 1.40.0). Conserved and species-specific genes were determined based on HGNC symbol and the intersection matrix obtained by `getMatrix` in GeneOverlap. Gene and disease ontology enrichments were calculated using clusterProfiler.

Final Seurat objects were prepared for display in an interactive webapp using the ShinyCell R Package¹³⁹.

Subclustering of major cell types

For major cell types labelled as mesenchymal and ectodermal, further subclustering was first performed on mouse data. For each major cell type, normalization, scaling with regressed cell cycle impacts, harmonization, and subclusters were identified using the same procedure as described above. Marker genes were identified for each cluster and functional enrichments were determined using clusterProfiler. Annotations for each cluster were manually assigned based on those originally described^{13–16}. Mouse cell subtype assignments were also confirmed with ToppGene¹⁴⁰ using the `toppFun` function from the scToppR package¹⁴¹. Mouse main and subtype annotations were further confirmed by projection on mouse E15.5 spatial transcriptomics data²³ (GSE245469). Links between snRNA and spatial data were determined using FindTransferAnchors and transferred using TransferData in Seurat.

Following annotation of subtypes and for comparison with human data, an intermediate data set was created where mouse genes were

reduced and converted to those to those with one-to-one orthology with human genes using the orthogene R package (v. 1.10.0) as described above. The intermediate dataset was used to transfer mouse subtype annotations to human subtypes by first identifying shared features across clusters using FindTransferAnchors in Seurat with log normalization and canonical correlation analysis (cca) or reciprocal PCA projection (rpca). Predicted subtype labels were transferred to human subtypes using TransferData in Seurat and further confirmed with a confusion matrix. As an orthogonal approach to label transferring in Seurat, we used SAMap⁷² (v. 1.0.15), an algorithm implemented in Python that uses self-assembling manifolds to map homologous cell types with shared expression programs across species. Briefly, we used the `map_genes` function within the SAMap package, which performs a BLAST search between species, with `t1` and `t2` both set to “nucl”, `tr1` set to the gencode v32 human fasta file, `tr2` set to the gencode v10 mouse fasta file, `n1` set to “hu” for human, and `n2` set to “mo” for mouse. This generates a “maps” directory with the transcriptome mapping BLAST results. Next, we used the `SAMAP`, `run`, and `get_mapping_scores` functions within the SAMap package to calculate alignment scores between cell types and subtypes across species. To finalize cell subtype annotations, we used a combinatorial approach considering the three annotation methods (label transfer by Seurat using cca, label transfer by Seurat using rpca, and label mapping using SAMap). Final annotations were confirmed using contingency matrices, significant marker gene overlap between human and mouse subtypes using the GeneOverlap R package¹³⁸, and functional enrichments using the `compareCluster` function in the clusterProfiler R package¹⁴².

The putative CNCC and early progenitor clusters were annotated by generating functional and/or regional module scores using marker genes from various publicly available databases and resources. To label the subclusters of the putative CNCC population, we generated a Schwann cell module score using the marker genes from Cao et al.²⁴. Additionally, we used the gene markers from various functionally specific, fate committed CNCCs in Soldatov et al.⁷⁰ and identified one cluster of mesenchymal-committed CNCCs and one cluster of sensory-committed CNCCs. The remaining putative CNCC subclusters were labeled according to Carnegie stage-specific biases. We randomly downsampled the number of cells to 200 cells per stage and quantified the percentage of cells in each cluster; those that were biased towards CS12 and CS13 were labeled as early (early.CNCC) and those that were biased towards CS13–C17 were labeled as intermediate (int.CNCC). When subclustering the early progenitor population, we generated module scores for neural progenitors using marker genes from the ScType database¹⁴³, neural stem cells using marker genes from Wang et al.²⁶, mesenchymal progenitors using markers from Wang et al.²⁶ and epithelial progenitors using markers from the human CellMarker2.0 database¹⁴⁴.

Marker genes for human subtypes were identified as performed for major cell types and functional enrichments were characterized with `compareCluster` in clusterProfiler. Final Seurat main objects and subtype objects were prepared for display in an interactive webapp using the ShinyCell R Package¹³⁹. Seurat objects were also converted to h5 files using the `SaveH5Seurat` and `Convert` functions in the SeuratDisk R package (v. 0.0.0.9021). These files are compatible with Python packages, including scanpy¹⁴⁵ and anndata¹⁴⁶ and were used for hosting at the Chan Zuckerberg Initiative CELLxGENE Discover resource¹⁴⁷.

Processing of spatial transcriptomics

Spatial transcriptomics data for two sagittal sections of a human CS13 embryo²⁵ were retrieved from <https://heoa.shinyapps.io/code/>. Slice 1 is more laterally localized, while Slice 2 is more medially localized. Thus, representative plots for ectoderm are shown in Slice 1 and for putative CNCC's and mesenchyme are shown in Slice 2.

Raw sequence count matrices were loaded using the Read10x command of Seurat¹³² and converted to HDF5 format. These counts were then combined with spot coordinates and section images using CreateSeuratObject. Data from both slices were merged and variable features were determined using Seurat. The percentage of mitochondrial reads was determined for each cell and was used to transform all data in the merged object using SCTransform¹⁴⁸ from Seurat. Data was clustered using UMAP and plotted, which revealed a strong batch effect between the two spatial objects. Data was further normalized using Harmony (v. 1.2.3), projected using UMAP, and clustered with a resolution of 0.8. Marker genes for the 22 clusters were identified using FindAllMarkers in Seurat. The top 100 marker genes for each cluster ($p < 0.05$, ranked by log2fold change versus all other clusters) were then analyzed for gene and disease ontology enrichments using compareCluster in clusterProfiler R package (v. 4.12.6). Enrichments and spatial localization were compared to previous annotations by Xu et al.²⁵ and labelled accordingly. The top 100 marker genes from each of the subclusters identified in human craniofacial data were used to calculate module scores across the merged spatial object and plotted using SpatialFeaturePlot in Seurat.

E9.5–16.5 spatial transcriptomics datasets were retrieved from the MOSTA resource⁷¹ and loaded into Seurat as for human data above. For representative images in the supplement, we chose to use the mouse E11.5 spatial transcriptomics data as it is most morphologically similar to CS13 human embryos. For the E9.5–E16.5 combined stages object, we converted the mouse genes to those with one-to-one orthology with human genes using the orthogene R package (v. 1.10.0). Following, we used FindTransferAnchors in Seurat with log normalization and canonical correlation analysis (cca) to transfer the human main cell type and subtype annotations to the mouse spatial transcriptomics dataset using only the highly variable genes identified using FindVariableFeatures in Seurat. Predicted labels were then transferred to mouse spatial transcriptomics data using TransferData in Seurat. For each E11.5 section, module scores of top 100 marker genes for each main cluster or subcluster were calculated. Gene spatial feature plots for selected genes and module scores were then generated with Seurat.

Facial variation and congenital abnormality GWAS enrichments

We retrieved summary statistics for facial variation^{87,88} and all congenital abnormality GWAS summary statistics from FinnGenn⁹². Raw summary stats were further processed and standardized with hg38 coordinates with MungeSumstats R package¹⁴⁹ (v. 1.10.1). Variants were mapped to genes ± 100 kb using MAGMA¹⁵⁰. Frontal facial measures and FinnGenn data were processed based on 1000 Genome European population, while profile facial measures were processed with the 1000 Genome Middle/South American population all obtained from the MAGMA website (<https://cncr.nl/research/magma/>). We converted the Seurat snRNA-seq expression data to a CellTypeData set with the Expression Weighted Celltype Enrichment (EWCE) R package¹⁵¹ (v. 1.12.0) and then assessed each study trait for a linear positive correlation of cell type gene expression specificity and gene-level genetic associations using MAGMA.Celltyping¹⁵² (v. 2.0.14). Plots were generated using tidyheatmaps¹⁵³ (v. 0.2.1) and ggplot2 (v. 3.5.1) in R.

Spatial GWAS enrichments

The human CS13 spatial transcriptomics Seurat object was subset to include only Slice 1 and then converted to h5ad format using the Convert function from the SeuratDisk package in R. GWAS summary statistics of facial variation and congenital abnormalities from FinnGenn were then mapped to the human CS13 spatial transcriptomics data using gsMap⁹³. The format_sumstats function was first used to convert the FinnGenn GWAS summary statistics to a format compatible by gsMap. A sumstats configuration file was then generated with information for all GWAS traits analyzed and then gsmap quick_run was run with default parameters. The gsMap run_cauchy_combination and

run_report functions were run for each GWAS trait to visualize the results. Genetic spatial mapping and gene specificity score plots were generated using Seurat's SpatialFeaturePlot function in R.

Gene list enrichments per cell type

The CellTypeData-formatted human craniofacial snRNA-seq objects were generated using generate_celltype_data in EWCE (v. 1.12.0). Mean and specificity metrics for several marker genes (*SOX10*, *TP63*, and *MSX1*) were inspected across main cell types and subtypes using plot_ctd in EWCE. Gene lists were compiled from multiple resources, including gnomAD (v4.1), CleftGeneDB, prioritized genes, and black module from craniofacial WGCNA¹³, and genes affected by de novo variation in orofacial clefting probands^{32,97}. For Neanderthal introgressed regions and HARs coordinates were obtained from respective publications^{154,155} and assigned single nearest gene using rGREAT (v. 2.6.0)¹⁵⁶ with “oneClosest” association rule. Each gene list was then tested for linear association using bootstrap enrichment test in EWCE (reps = 10,000; geneSizeControl = TRUE). As an additional negative control, we generated random gene lists based on the mean number of all the lists tested here. Results from all gene lists were then merged, corrected for the total number of gene lists and cell types tested using the Benjamini–Hochberg approach, and then plotted using the pheatmap package¹⁵⁷ (v. 1.0.12) in R.

Phenotype-cell type association tests

To map the relationships between cell types and phenotypes, we ran pairwise association tests between all combinations of cell types in our snRNA-seq-derived CellTypeData and phenotypes across the Human Phenotype Ontology (HPO; release v2025-08-11)⁹⁰ using the run_phenomix function from MSTExplorer (v. 1.0.9). In contrast to the gene list-based approaches (e.g., EWCE) this function reframes the problem as a series of generalized linear regressions by leveraging continuous scores that summarize the current strength of evidence for a causal relationship between each gene-phenotype pair (using additional data from the Gene Curation Coalition)^{89,158}. The continuous nature of these data allows us to more accurately capture phenotype-cell type relationships, especially for phenotypes with large gene lists where only some genes have strong causal evidence for the phenotype. The gene signature vectors for each phenotype were previously merged and shared as a single precomputed gene (5180 unique gene symbols) $\times$ phenotype (11,606 unique HPO phenotypes) association matrix. Next, a series of linear regressions tests were performed between the gene specificity vectors of each cell subtype ($n = 89$ vectors) and the gene association vectors of each phenotype ($n = 11,606$ vectors). Finally, multiple-testing correction was applied using Benjamini–Hochberg False Discovery Rate¹⁰⁵ (at FDR < 5% significance).

For the purposes of summarization and visualization, the number of significantly associated phenotypes per cell type were then computed within each major HPO branch (Fig. 7c). Here, we define HPO branches as groups of related phenotypes that can be labeled according to their shared ancestral term, e.g., “Abnormality of the immune system”. Next, we sought to determine whether some cell types were disproportionately more often associated with phenotypes of a particular HPO branch. To accomplish this, we performed a series of proportion tests comparing the proportion of total phenotypes that a given cell type was significantly associated with within a target HPO branches relative to all other HPO branches. In practice, we computed 2×2 contingency tables (number of significant phenotype association vs. number of non-significant phenotype associations $\times$ target branch vs. non-target branches) for each cell type within each HPO branch, which were then used as inputs to the prop_test function within the rstatix R package¹⁵⁹ (v. 0.7.2). This test appropriately takes into account the different number of phenotypes across HPO branches. Only one-sided tests were performed to test whether the target HPO branch was greater than all other (non-target) branches (set with the

alternative = “greater” parameter). All proportion tests were then corrected for multiple testing at FDR < 5%.

Orofacial Clefting de novo variant analysis

We used the R package DenovolyzeR¹⁰³ (v. 0.2.0) to test enrichment of de novo variants (DNs) in a dataset of OFC case-parent trios. Enrichment is calculated by comparing the expected number of variants, as determined by mutation models described by Samocha et al.¹⁰⁴, to the observed number of variants in a given gene or group of genes using the “DenovolyzeByClass” and “includeGenes” functions. Using our dataset of 2031 DN in 1171 genes identified in 1676 trios with OFCs, we first compared this list of genes to those with calculated mutational rates in the R package “DenovolyzeR” using the “view-ProbabilityTable()” function. Broken down by subtype, there were 1180 cleft lip with or without cleft palate (CL/P; 226 cleft lip (CL), 966 cleft lip and palate (CLP)), and 484 cleft palate (CP) trios. We then tested enrichment of all OFC trios and by subtype within the top 20% of genes by log2FC derived from single nucleus RNA sequencing of human craniofacial tissue at CS20.

Statistics and reproducibility. Detailed statistical analyses are described in the “Methods” section above. Biological replicates for the human embryonic face samples were collected in the following way: four at CS12, six at CS13, three at CS14, three at CS16, three at CS17, and five at CS20 for single nucleus multiomics. Biological replicates of pooled mouse embryonic face samples were collected in the following way: three at E10.5, three at E11.5, and three at E12.5 for single nucleus multiomics. Appropriate negative controls were used throughout the study for GWAS enrichment across main cell types and subtypes. The investigators were not blinded to allocation during experiments and outcome assessment. No statistical method was used to pre-determine sample size. No data that passed quality control criteria for experiments were excluded from the analyses. The experiments were not randomized. Unless otherwise stated, default parameter settings were employed for any software tool that was used in the analyses. Whenever a *p*-value is reported in the text, the statistical test is also indicated. All statistics were calculated, and plots were generated using R v4.4.0.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw human and mouse data generated in this study have been deposited in the FaceBase database under accessions [8D-2JQO¹⁶⁰](#) and [7N-54BY¹⁶¹](#), respectively. Raw data from human samples is controlled access. Access request procedures can be found on [Facebase](#). The processed cellranger ARC gene expression outputs for both human and mouse are available on [Zenodo¹⁶²](#). Source data of all plots are available in Seurat R objects at the [Zenodo repository¹⁶²](#). An interactive browser of these data can be found at https://cotneyshiny.research.chop.edu/shiny-apps/craniofacial_all_snRNA_v2/. The data can also be explored on the Chan Zuckerberg CellxGene Discover Database (<https://cellxgene.cziscience.com/collections/690fb28b-e364-4d3d-8934-bae50ae93857>).

Code availability

Code for analysis and generation of figures can be found on GitHub (https://github.com/cotneylab/craniofacial_snRNA)¹⁶³.

References

- Minoux, M. & Rijli, F. M. Molecular mechanisms of cranial neural crest cell migration and patterning in craniofacial development. *Development* **137**, 2605–2621 (2010).
- Cordero, D. R. et al. Cranial neural crest cells on the move: their roles in craniofacial development. *Am. J. Med. Genet. Part A* **155**, 270–279 (2011).
- Tang, W. & Bronner, M. E. Neural crest lineage analysis: from past to future trajectory. *Development* **147**, dev193193 (2020).
- Guo, J. et al. Variation and signatures of selection on the human face. *J. Hum. Evol.* **75**, 143–152 (2014).
- Mitteroecker, P., Gunz, P., Bernhard, M., Schaefer, K. & Bookstein, F. L. Comparison of cranial ontogenetic trajectories among great apes and humans. *J. Hum. Evol.* **46**, 679–697 (2004).
- Naqvi, S. et al. Decoding the human face: progress and challenges in understanding the genetics of craniofacial morphology. *Annu. Rev. Genomics Hum. Genet.* **23**, 383–412 (2022).
- Smith, D. W. Recognizable patterns of human malformation: genetic, embryologic, and clinical aspects. *Major Probl. Clin. Pediatr.* **7**, 1–368 (1970).
- Leslie, E. J. & Marazita, M. L. Genetics of cleft lip and cleft palate. *Am. J. Med. Genet. Part C, Semin. Med. Genet.* **163C**, 246–258 (2013).
- Mc Goldrick, N. et al. A multi-program analysis of cleft lip with cleft palate prevalence and mortality using data from 22 International Clearinghouse for Birth Defects Surveillance and Research programs, 1974–2014. *Birth Defects Res.* **115**, 980–997 (2023).
- Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J. & Shaw, W. C. Cleft lip and palate. *Lancet* **374**, 1773–1785 (2009).
- Mai, C. T. et al. National population-based estimates for major birth defects, 2010–2014. *Birth Defects Res.* **111**, 1420–1435 (2019).
- Li, H., Jones, K. L., Hooper, J. E. & Williams, T. The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. *Development* **146**, dev174888 (2019).
- Yankee, T. N. et al. Integrative analysis of transcriptome dynamics during human craniofacial development identifies candidate disease genes. *Nat. Commun.* **14**, 4623 (2023).
- Rajderkar, S. S. et al. Dynamic enhancer landscapes in human craniofacial development. *Nat. Commun.* **15**, 2030 (2024).
- Sun, J. et al. Single-cell RNA-Seq reveals transcriptional regulatory networks directing the development of mouse maxillary prominence. *J. Genet. Genomics* **50**, 676–687 (2023).
- Han, X. et al. Runx2-Twist1 interaction coordinates cranial neural crest guidance of soft palate myogenesis. *Elife* **10**, e62387 (2021).
- Mestas, J. & Hughes, C. C. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
- Martik, M. L. & Bronner, M. E. Riding the crest to get a head: neural crest evolution in vertebrates. *Nat. Rev. Neurosci.* **22**, 616–626 (2021).
- York, J. R., Yuan, T. & McCauley, D. W. Evolutionary and developmental associations of neural crest and placodes in the vertebrate head: insights from jawless vertebrates. *Front. Physiol.* **11**, 986 (2020).
- Donoghue, P. C. J., Graham, A. & Kelsh, R. N. The origin and evolution of the neural crest. *BioEssays* **30**, 530–541 (2008).
- Selleri, L. & Rijli, F. M. Shaping faces: genetic and epigenetic control of craniofacial morphogenesis. *Nat. Rev. Genet.* **24**, 610–626 (2023).
- Martínez-Abadías, N. et al. The developmental basis of quantitative craniofacial variation in humans and mice. *Evolut. Biol.* **39**, 554–567 (2012).
- Pina, J. O. et al. Multimodal spatiotemporal transcriptomic resolution of embryonic palate osteogenesis. *Nat. Commun.* **14**, 5687 (2023).
- Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
- Xu, Y. et al. A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nat. Cell Biol.* **25**, 604–615 (2023).

26. Wang, C. et al. Single-cell RNA sequencing analysis of human embryos from the late Carnegie to fetal development. *Cell Biosci.* **14**, 118 (2024).
27. Zeng, B. et al. The single-cell and spatial transcriptional landscape of human gastrulation and early brain development. *Cell Stem Cell* **30**, 851–866.e7 (2023).
28. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J. P. & Cotney, J. High-resolution epigenomic atlas of human embryonic craniofacial development. *Cell Rep.* **23**, 1581–1597 (2018).
29. Vieille-Grosjean, I., Hunt, P., Gulisano, M., Boncinelli, E. & Thorgood, P. Branchial HOX gene expression and human craniofacial development. *Dev. Biol.* **183**, 49–60 (1997).
30. Cai, J. et al. Gene expression in pharyngeal arch 1 during human embryonic development. *Hum. Mol. Genet.* **14**, 903–912 (2005).
31. Samuels, B. D. et al. FaceBase 3: analytical tools and FAIR resources for craniofacial and dental research. *Development* **147**, dev191213 (2020).
32. Bishop, M. R. et al. Genome-wide enrichment of de novo coding mutations in orofacial cleft trios. *Am. J. Hum. Genet.* **107**, 124–136 (2020).
33. Schoenwolf, G. C., Bleyl, S. B., Brauer, P. R. & Francis-West, P. *Larsen's Human Embryology* (Elsevier, Philadelphia, PA, 2021).
34. Jirasek, J. E. *An Atlas of Human Prenatal Developmental Mechanics: Anatomy and Staging*, 312 (CRC Press, London, 2004).
35. CZI Cell Science Program, et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
36. Leland, M., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at <https://arxiv.org/abs/1802.03426> (2020).
37. Dobrev, G. et al. SATB2 is a multifunctional determinant of craniofacial patterning and osteoblast differentiation. *Cell* **125**, 971–986 (2006).
38. Huang, X. et al. SATB2: a versatile transcriptional regulator of craniofacial and skeleton development, neurogenesis and tumorigenesis, and its applications in regenerative medicine. *Genes Dis.* **9**, 95–107 (2022).
39. Jin, Y. R., Turcotte, T. J., Crocker, A. L., Han, X. H. & Yoon, J. K. The canonical Wnt signaling activator, R-spondin2, regulates craniofacial patterning and morphogenesis within the branchial arch through ectodermal-mesenchymal interaction. *Dev. Biol.* **352**, 1–13 (2011).
40. Alhazmi, N. et al. Synergistic roles of Wnt modulators R-spondin2 and R-spondin3 in craniofacial morphogenesis and dental development. *Sci. Rep.* **11**, 5871 (2021).
41. Nikolopoulou, E. et al. Spinal neural tube closure depends on regulation of surface ectoderm identity and biomechanics by Grhl2. *Nat. Commun.* **10**, 2487 (2019).
42. Knowles, W. J., Bologna, M. L., Chasis, J. A., Marchesi, S. L. & Marchesi, V. T. Common structural polymorphisms in human erythrocyte spectrin. *J. Clin. Invest.* **73**, 973–979 (1984).
43. Bennett, V. & Stenbuck, P. J. The membrane attachment protein for spectrin is associated with band 3 in human erythrocyte membranes. *Nature* **280**, 468–473 (1979).
44. Huang, C. H. The human Rh50 glycoprotein gene. Structural organization and associated splicing defect resulting in Rh(null) disease. *J. Biol. Chem.* **273**, 2207–2213 (1998).
45. Vallese, F. et al. Architecture of the human erythrocyte ankyrin-1 complex. *Nat. Struct. Mol. Biol.* **29**, 706–718 (2022).
46. Kendall, R. L., Wang, G. & Thomas, K. A. Identification of a natural soluble form of the vascular endothelial growth factor receptor, FLT-1, and its heterodimerization with KDR. *Biochem. Biophys. Res. Commun.* **226**, 324–328 (1996).
47. Kabir, A. U. et al. Dual role of endothelial Myc1 in tumor angiogenesis and tumor immunity. *Sci. Transl. Med.* **13**, eabb6731 (2021).
48. Ferrero, G. et al. The macrophage-expressed gene (mpeg) 1 identifies a subpopulation of B cells in the adult zebrafish. *J. Leukoc. Biol.* **107**, 431–443 (2020).
49. Engel, P. et al. The B7-2 (B70) costimulatory molecule expressed by monocytes and activated B lymphocytes is the CD86 differentiation antigen. *Blood* **84**, 1402–1407 (1994).
50. Justement, L. B., Campbell, K. S., Chien, N. C. & Cambier, J. C. Regulation of B cell antigen receptor signal transduction and phosphorylation by CD45. *Science* **252**, 1839–1842 (1991).
51. Wang, M. H. et al. Identification of the ron gene product as the receptor for the human macrophage stimulating protein. *Science* **266**, 117–119 (1994).
52. Fabrick, B. O. et al. The macrophage scavenger receptor CD163 functions as an innate immune sensor for bacteria. *Blood* **113**, 887–892 (2009).
53. Weintraub, H. et al. The myoD gene family: nodal point during specification of the muscle cell lineage. *Science* **251**, 761–766 (1991).
54. Seidel, U. & Arnold, H. H. Identification of the functional promoter regions in the human gene encoding the myosin alkali light chains MLC1 and MLC3 of fast skeletal muscle. *J. Biol. Chem.* **264**, 16109–16117 (1989).
55. Karsch-Mizrachi, I., Travis, M., Blau, H. & Leinwand, L. A. Expression and DNA sequence analysis of a human embryonic skeletal muscle myosin heavy chain gene. *Nucleic Acids Res.* **17**, 6167–6179 (1989).
56. Sasai, N., Mizuseki, K. & Sasai, Y. Requirement of FoxD3-class signaling for neural crest determination in *Xenopus*. *Development* **128**, 2525–2536 (2001).
57. Kos, R., Reedy, M. V., Johnson, R. L. & Erickson, C. A. The winged-helix transcription factor FoxD3 is important for establishing the neural crest lineage and repressing melanogenesis in avian embryos. *Development* **128**, 1467–1479 (2001).
58. Lukoseviciute, M. et al. From pioneer to repressor: bimodal foxd3 activity dynamically remodels neural crest regulatory landscape *In Vivo*. *Dev. Cell* **47**, 608–628.e6 (2018).
59. Simões-Costa, M. S., McKeown, S. J., Tan-Cabugao, J., Sauka-Spengler, T. & Bronner, M. E. Dynamic and differential regulation of stem cell factor FoxD3 in the Neural Crest Is Encrypted in the Genome. *PLoS Genet.* **8**, e1003142 (2012).
60. Mandalos, N. et al. Sox2 acts as a rheostat of epithelial to mesenchymal transition during neural crest development. *Front. Physiol.* **5**, 345 (2014).
61. Mandalos, N. P., Dimou, A., Gavala, M. A., Lambraki, E. & Remboutsika, E. Craniofacial development is fine-tuned by Sox2. *Genes* **14**, 380 (2023).
62. Thakurela, S. et al. Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov.* **2**, 15045 (2016).
63. Chen, C. et al. The transcription factor POU3F2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. *Sci. Transl. Med.* **10**, eaat8178 (2018).
64. Simões-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development* **142**, 242–257 (2015).
65. Feng, Z. et al. hReg-CNCC reconstructs a regulatory network in human cranial neural crest cells and annotates variants in a developmental context. *Commun. Biol.* **4**, 442 (2021).
66. Thomas, S. et al. Human neural crest cells display molecular and phenotypic hallmarks of stem cells. *Hum. Mol. Genet.* **17**, 3411–3425 (2008).

67. Betters, E., Liu, Y., Kjaeldgaard, A., Sundstrom, E. & Garcia-Castro, M. I. Analysis of early human neural crest development. *Dev. Biol.* **344**, 578–592 (2010).
68. O’Rahilly, R. & Müller, F. The development of the neural crest in the human. *J. Anat.* **211**, 335–351 (2007).
69. Kastri, M. E. et al. Schwann cell precursors represent a neural crest-like state with biased multipotency. *EMBO J.* **41**, e108780 (2022).
70. Soldatov, R. et al. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
71. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
72. Tarashansky, A. J. et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**, e66747 (2021).
73. Aoto, K., Nishimura, T., Eto, K. & Motoyama, J. Mouse GLI3 regulates Fgf8 expression and apoptosis in the developing neural tube, face, and limb bud. *Dev. Biol.* **251**, 320–332 (2002).
74. Baker, J. L., Wood, B., Karpinski, B. A., LaMantia, A. S. & Maynard, T. M. Testicular receptor 2, Nr2c1, is associated with stem cells in the developing olfactory epithelium and other cranial sensory and skeletal structures. *Gene Expr. Patterns* **20**, 71–79 (2016).
75. Mansouri, A., Hallonet, M. & Gruss, P. Pax genes and their roles in cell differentiation and development. *Curr. Opin. Cell Biol.* **8**, 851–857 (1996).
76. Nakashima, K. et al. The novel zinc finger-containing transcription factor osterix is required for osteoblast differentiation and bone formation. *Cell* **108**, 17–29 (2002).
77. Hojo, H., Ohba, S., He, X., Lai, L. P. & McMahon, A. P. Sp7/osterix is restricted to bone-forming vertebrates where it acts as a Dlx Co-factor in osteoblast specification. *Dev. Cell* **37**, 238–253 (2016).
78. Sur, A. et al. Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development. *Dev. Cell* **58**, 3028–3047.e12 (2023).
79. Van Otterloo, E. et al. AP-2 α and AP-2 β cooperatively function in the craniofacial surface ectoderm to regulate chromatin and gene expression dynamics during facial development. *Elife* **11**, e70511 (2022).
80. de la Garza, G. et al. Interferon regulatory factor 6 promotes differentiation of the periderm by activating expression of Grainyhead-like 3. *J. Invest. Dermatol.* **133**, 68–77 (2013).
81. Wang, Y. et al. Otoconin-90, the mammalian otoconial matrix protein, contains two domains of homology to secretory phospholipase A2. *Proc. Natl. Acad. Sci. USA* **95**, 15345–15350 (1998).
82. Zhao, X., Yang, H., Yamoah, E. N. & Lundberg, Y. W. Gene targeting reveals the role of Oc90 as the essential organizer of the otoconial organic matrix. *Dev. Biol.* **304**, 508–524 (2007).
83. De Felice, M. et al. A mouse model for hereditary thyroid dysgenesis and cleft palate. *Nat. Genet.* **19**, 395–398 (1998).
84. Sheng, H. Z. et al. Specification of pituitary cell lineages by the LIM homeobox gene Lhx3. *Science* **272**, 1004–1007 (1996).
85. Raetzman, L. T., Ward, R. & Camper, S. A. Lhx4 and Prop1 are required for cell survival and expansion of the pituitary primordia. *Development* **129**, 4229–4239 (2002).
86. Piña, J. O. et al. Spatial Multi-omics Reveals the Role of the Wnt Modulator, Dkk2, in Palatogenesis. *J. Dent. Res.* **103**, 1412–1420 (2024).
87. Xiong, Z. et al. Novel genetic loci affecting facial shape variation in humans. *eLife* **8**, e49898 (2019).
88. Bonfante, B. et al. A GWAS in Latin Americans identifies novel face shape loci, implicating VPS13B and a Denisovan introgressed region in facial variation. *Sci. Adv.* **7**, eabc6160 (2021).
89. Murphy, K. B. et al. Identification of cell type-specific gene targets underlying thousands of rare diseases and subtraits. Preprint at <https://www.medrxiv.org/content/10.1101/2023.02.13.23285820v1> (2023).
90. Gargano, M. A. et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2023).
91. Schilder, B. M. et al. Cell type-specific contextualisation of the human phenotype: towards the systematic treatment of all rare diseases. Preprint at <https://www.medrxiv.org/content/10.1101/2023.02.13.23285820v2> (2025).
92. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
93. Song, L., Chen, W., Hou, J., Guo, M. & Yang, J. Spatially resolved mapping of cells associated with human complex traits. *Nature* **641**, 932–941 (2025).
94. Xu, H. et al. CleftGeneDB: a resource for annotating genes associated with cleft lip and cleft palate. *Sci. Bull.* **66**, 2340–2342 (2021).
95. Mukhopadhyay, N. et al. Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21. *Hum. Genet.* **139**, 215–226 (2020).
96. Curtis, S. W. et al. Rare variant modifier analysis identifies variants in SEC24D associated with orofacial cleft subtypes. *Hum. Genet.* **142**, 1531–1541 (2023).
97. Robinson, K. et al. Trio-based GWAS identifies novel associations and subtype-specific risk factors for cleft palate. *HGG Adv.* **4**, 100234 (2023).
98. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
99. Doan, R. N. et al. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell* **167**, 1–27 (2016).
100. Levchenko, A., Kanapin, A., Samsonova, A. & Gainetdinov, R. R. Human accelerated regions and other human-specific sequence variations in the context of evolution and their relevance for brain development. *Genome Biol. Evol.* **10**, 166–188 (2018).
101. Driessens, S. L. W. et al. Genes associated with cognitive ability and HAR show overlapping expression patterns in human cortical neuron types. *Nat. Commun.* **14**, 4188 (2023).
102. Prabhakar, S. et al. Human-specific gain of function in a developmental enhancer. *Science* **321**, 1346–1350 (2008).
103. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15 (2015).
104. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
105. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B-Methodol.* **57**, 289–300 (1995).
106. Prescott, Sara, L. et al. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).
107. Ye, Q., Bhojwani, A. & Hu, J. K. Understanding the development of oral epithelial organs through single cell transcriptomic analysis. *Development* **149**, dev200539 (2022).
108. Weinberg, S. M., Cornell, R. & Leslie, E. J. Craniofacial genetics: where have we been and where are we going? *PLoS Genet.* **14**, e1007438 (2018).
109. Claes, P. et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* **50**, 414–423 (2018).
110. White, J. D. et al. Insights into the genetic architecture of the human face. *Nat. Genet.* **53**, 45–53 (2021).
111. Dixon, M. J., Marazita, M. L., Beaty, T. H. & Murray, J. C. Cleft lip and palate: understanding genetic and environmental influences. *Nat. Rev. Genet.* **12**, 167–178 (2011).

112. Marazita, M. L. The evolution of human genetic studies of cleft lip and cleft palate. *Annu. Rev. Genomics Hum. Genet.* **13**, 263–283 (2012).
113. Rahimov, F. et al. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat. Genet.* **40**, 1341–1347 (2008).
114. Indencleef, K. et al. Six NSCL/P loci show associations with normal-range craniofacial variation. *Front. Genet.* **9**, 502 (2018).
115. Indencleef, K. et al. The intersection of the genetic architectures of orofacial clefts and normal facial variation. *Front. Genet.* **12**, 626403 (2021).
116. Howe, L. J. et al. Investigating the shared genetics of non-syndromic cleft lip/palate and facial morphology. *PLoS Genet.* **14**, e1007501 (2018).
117. Weinberg, S. M. What's shape got to do with it? Examining the relationship between facial shape and orofacial clefting. *Front. Genet.* **13**, 891502 (2022).
118. Tamarin, A. & Boyde, A. Facial and visceral arch development in the mouse embryo: a study by scanning electron microscopy. *J. Anat.* **124**, 563–580 (1977).
119. Depew, M. J. & Compagnucci, C. Tweaking the hinge and caps: testing a model of the organization of jaws. *J. Exp. Zool. B Mol. Dev. Evol.* **310**, 315–335 (2008).
120. Losa, M. et al. Face morphogenesis is promoted by Pbx-dependent EMT via regulation of Snail1 during frontonasal prominence fusion. *Development* **145**, dev157628 (2018).
121. Jiang, R., Bush, J. O. & Lidral, A. C. Development of the upper lip: morphogenetic and molecular mechanisms. *Dev. Dyn.* **235**, 1152–1166 (2006).
122. Wang, K. H. et al. Evaluation and integration of disparate classification systems for clefts of the lip. *Front. Physiol.* **5**, 163 (2014).
123. Lin, H. & Grosschedl, R. Failure of B-cell differentiation in mice lacking the transcription factor EBF. *Nature* **376**, 263–267 (1995).
124. Jin, S. et al. Ebf factors and MyoD cooperate to regulate muscle relaxation via Atp2a1. *Nat. Commun.* **5**, 3793 (2014).
125. Garcia-Dominguez, M., Poquet, C., Garel, S. & Charnay, P. Ebf gene function is required for coupling neuronal differentiation and cell cycle exit. *Development* **130**, 6013–6025 (2003).
126. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
127. Jones, D. L. & Wagers, A. J. No place like home: anatomy and function of the stem cell niche. *Nat. Rev. Mol. Cell Biol.* **9**, 11–21 (2008).
128. Zardo, G. et al. Integrated genomic and epigenomic analyses pinpoint biallelic gene inactivation in tumors. *Nat. Genet.* **32**, 453–458 (2002).
129. Liao, D. Emerging roles of the EBF family of transcription factors in tumor suppression. *Mol. Cancer Res.* **7**, 1893–1901 (2009).
130. Karl, T. *The House Mouse: Atlas of Embryonic Development* (Springer Science+Business Media, 1989).
131. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
132. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
133. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
134. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* **24**, 584–594 (2021).
135. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
136. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
137. Schilder, B. M., Murphy, A. E. & Skene, N. G. orthogene: a Bioconductor package to easily map genes within and across hundreds of species. *bioRxiv* <https://doi.org/10.64898/2026.01.17.700094> (2026).
138. Shen, L. & Sinai, I. IsoMaM. GeneOverlap: Test and visualize gene overlaps. <https://doi.org/10.18129/B9.bioc.GeneOverlap>.
139. Ouyang, J. F., Kamaraj, U. S., Cao, E. Y. & Rackham, O. J. L. ShinyCell: simple and sharable visualization of single-cell gene expression data. *Bioinformatics* **37**, 3374–3376 (2021).
140. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
141. Granger, B. & Berto, S. scToppR: a coding-friendly R interface to ToppGene. *Bioinformatics* **40**, btac582 (2024).
142. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
143. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).
144. Hu, C. et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).
145. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
146. Virshup, I. et al. anndata: Access and store annotated data matrices. *J. Open Sour. Softw.* **9**, 4371 (2024).
147. Program, C. Z. I. C. S. et al. CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).
148. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
149. Murphy, A. E., Schilder, B. M. & Skene, N. G. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. *Bioinformatics* **37**, 4593–4596 (2021).
150. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
151. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
152. Skene, N. G. et al. Genetic identification of brain cell types underlying Schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
153. Engler, J. tidyheatmaps: Heatmaps from Tidy Data. <https://doi.org/10.32614/CRAN.package.tidyheatmaps> (2024).
154. McArthur, E., Rinker, D. C. & Capra, J. A. Quantifying the contribution of Neanderthal introgression to the heritability of complex traits. *Nat. Commun.* **12**, 4481 (2021).
155. Keough, K. C. et al. Three-dimensional genome rewiring in loci with human accelerated regions. *Science* **380**, eabm1696 (2023).
156. Gu, Z. & Hübschmann, D. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics* **39**, btac745 (2023).
157. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0.13. <https://github.com/raivokolde/pheatmap> (2025).
158. DiStefano, M. T. et al. The Gene Curation Coalition: a global effort to harmonize gene-disease evidence resources. *Genet. Med.* **24**, 1732–1742 (2022).
159. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.2, <https://rpkgs.datanovia.com/rstatix/> (2023).

160. Manchel, A. & Cotney, J. *Gene Expression Patterns of the Developing Human Face at Single Cell Resolution (snRNA-seq)*. <https://doi.org/10.25550/8D-2JQ> (FaceBase Consortium, 2026).
161. Manchel, A. & Cotney, J. *Gene Expression Patterns of the Developing Mouse Face at Single Cell Resolution (snRNA-seq)*. <https://doi.org/10.25550/7N-54BY> (FaceBase Consortium, 2026).
162. Cotney, J., Farah, N., Manchel, A. & Wentworth, E. Craniofacial multiome data (Version v2) [Data set]. Zenodo <https://doi.org/10.5281/zenodo.14675171> (2025).
163. Cotney, J. Cotneylab/craniofacial_snRNA. Zenodo <https://doi.org/10.5281/zenodo.18246448> (2026).

Acknowledgements

We would like to thank members of the UConn/JAXGM Single Cell Genomics Core for help with standardizing single-cell isolation techniques and preparing sequencing libraries. We would also like to thank members of the UConn Computational Biology Core and High-Performance Computing Facility for assistance with package installation and software/hardware support. We are grateful to Dr. Peter Tran, Dr. Sungryong Oh, and Pooja Sonawane for constructive feedback and copyediting. This work was funded by grants from the National Institutes of Health to E.J.L.C. (R01-DE030342, X01-HG010835, X01-HD100701, X01-HL132363) and J.C. (NIDCR 1R01DE028945, NIDCR 1R03DE028588, and NIGMS 5R35GM119465).

Author contributions

Conceptualization: J.C. Investigation: N.F., A.M., E.W.W., and J.C. Formal analysis: N.F., A.M., E.W.W., B.M.S., K.R., S.W.C., J.C. Writing—original draft: J.C. Writing—review and editing: N.F., A.M., E.W.W., B.S., K.R., S.W.C., N.G.S., E.J.L.C., J.C. Funding acquisition: J.C. Supervision: J.C., E.J.L.C., S.G.K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-70232-6>.

Correspondence and requests for materials should be addressed to Justin Cotney.

Peer review information *Nature Communications* thanks Haoyang Li and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026