Combining genetic and single-cell expression data to expand the understanding of orofacial clefting

Doctoral thesis
to obtain a doctorate (PhD)
from the Faculty of Medicine
of the University of Bonn

Anna Siewert
from Dinslaken, Germany
2025

Written with authorization of

the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Kerstin U. Ludwig

Second reviewer: Prof. Dr. Jan Hasenauer

Day of oral examination: March 20th, 2025

From the Institute of Human Genetics

To my dad,

who taught me that life should be lived to the fullest and that laughter is the best medicine. And while life without you by my side is highly illogical, you have been, and always shall be, my friend 🖖

Table of contents

1 List of abbreviations

BFAR	Bifunctional apoptosis regulator
CNCC	Cranial neural crest cells
CTNND1	Catenin delta 1
ESRP1	Epithelial splicing regulatory protein 1
FGFR1	Fibroblast growth factor receptor 1
FOXE1	Forkhead box E1
GWAS	Genome-wide association study
HAND2	Heart and neural crest derivatives expressed 2
HYAL2	Hyaluronidase 2
IRF6	Interferon regulatory factor 6
KRT18	Keratin 18
KRT8	Keratin 8
MSC	Musculin
MSX1	Homeobox protein MSX-1
NC	Neural crest
nsCL/P	Non-syndromic cleft lip with or without cleft palate
OFC	Orofacial clefting
PRTG	Protogenin
RNA	Ribonucleic acid
scDRS	Single-cell disease relevance score
scRNA-seq	Single-cell RNA sequencing
TADs	Topologically associating domains
TFAP2A	Transcription factor AP-2 alpha
TP63	Tumor protein P63

2 Abstract

Molecular malfunctions during craniofacial development can lead to non-syndromic cleft lip with or without cleft palate (nsCL/P), a congenital malformation that affects the upper lip and palate. Treatment involves multidisciplinary approaches, including surgery and speech therapy. In addition to an increased risk of morbidities such as cancer, neurological and cardiovascular diseases, these interventions can represent a burden for those affected. NsCL/P occurs between the fourth and tenth week of embryonic development and is one of the most common birth defects with a prevalence of about 1 in 1,000 live births. The etiology is described as multifactorial and polygenic, including environmental and numerous genetic factors. To date, various genetic studies have identified over 45 genomic loci that are associated with a risk for nsCL/P. However, most of the genetic risk variants map to non-coding regions of the genome, and the target genes and affected cell types are mostly unknown. The aim of the present thesis was to examine gene expression patterns of nsCL/P candidate genes to identify cell types that are potentially involved in the development of nsCL/P. For this purpose, published single-cell RNA sequencing (scRNA-seq) data from embryonic mice were re-analyzed. These data were then used to study the expression patterns of candidate genes that were identified by genome-wide association studies (GWAS) and whole-genome sequencing data. In addition to confirming gene expression patterns that were described in previous functional studies, this revealed that most candidate genes are specifically expressed in either one of two groups of cell types, epithelial or mesenchymal cells. After scRNA-seq data from the heads of human embryos became first available, a systematic analysis of the joint gene expression of nsCL/P GWAS candidate genes in these data showed that epithelial cells and *HAND2*+ pharyngeal arches are associated with nsCL/P candidate genes, complementing the previous research in murine scRNA-seq data. Co-expression network analysis in these cell types was then used to identify potential interactions between candidate genes and to prioritize candidate genes by combining the results with the initial GWAS data. The results were consistent with previously described gene-gene interactions and revealed potential new interactions and candidate genes. Together, these analyses determined nsCL/P-associated cell types and demonstrated a novel strategy for the prioritization of candidate genes based on a combination of GWAS and scRNA-seq data.

3 Introduction and aims

Molecular malfunctions during craniofacial development can lead to orofacial clefting (OFC), which is a group of congenital facial malformations that mainly affect the growth and fusion of the upper lip and palate. OFC can be divided into several subtypes, depending on which facial structures are affected by clefting. Additionally, OFC subtypes can be divided into non-syndromic (or isolated) and syndromic forms. In syndromic forms, OFC is part of a more complex malformation syndrome with further developmental defects. A prominent example of a syndrome with an OFC phenotype is Van der Woude syndrome, which is characterized by bottom lip pits and either cleft lip with or without cleft palate or only a cleft palate (van der Woude, 1954; Kondo et al., 2002; Mangold et al., 2016).

Among the various OFC subtypes, non-syndromic cleft lip with or without cleft palate (nsCL/P) is the most prevalent. While the prevalence varies globally, the average prevalence is about 1 in 1,000 live births, making nsCL/P one of the most common birth defects (Mangold et al., 2011). It is characterized by clefting of the upper lip and, in some cases, additionally the palate (Fig. 1).

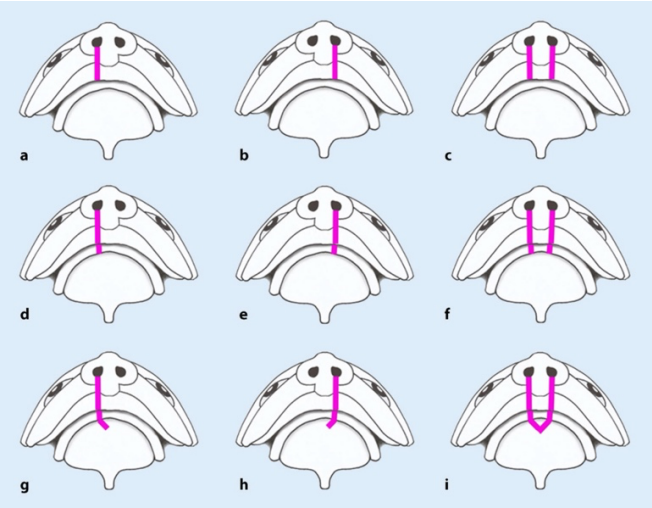


Figure 1: Frequent clefting patterns of facial structures (pink) for cleft lip with/without cleft palate; **a-c** of the upper lip; **d-f** of the upper lip and alveolar process; **g-i** of the upper lip, alveolar process and primary palate (figure and caption taken from Mangold, Kreiß and Nöthen, 2017)

Individuals with nsCL/P often require multidisciplinary interventions, which can include repeated corrective surgeries and speech therapy. Such interventions can be a

considerable financial and psychosocial burden for affected individuals and their families (Kousa and Schutte, 2016). Additionally, nsCL/P was found to be associated with an increased risk for morbidities such as cancer, neurological and cardiovascular diseases (Christensen et al., 2004; Dunkhase et al., 2016; Kousa and Schutte, 2016). Based on twin studies, the heritability of nsCL/P has been estimated to be approximately 90 %, indicating a strong genetic contribution (Grosen et al., 2011). Its etiology is described to be polygenic and multifactorial, with multiple genetic and environmental factors, including folic acid intake and alcohol consumption during pregnancy, being considered to play a role (Mangold et al., 2011).

The complex genetic architecture of nsCL/P is characterized by a certain degree of variability. There are monogenic forms of nsCL/P caused by rare variants that often have a high impact on risk. For example, studies have identified rare variants in the genes *MSX1* and *IRF6* to cause monogenic nsCL/P (Jezewski et al., 2003; Blanton et al., 2005; Pengelly et al., 2016). These monogenic forms often follow a Mendelian inheritance pattern within families (Gajdos et al., 2004). However, pedigrees of families with affected individuals are not always conclusive in terms of inheritance patterns. In most cases, nsCL/P is a polygenic condition that is likely to be caused by a multitude of common variants and does not follow a typical pattern of inheritance. In these cases, it is the cumulative effect of variants with small effects on risk that likely causes nsCL/P (Ludwig et al., 2017).

To date, numerous genetic studies have identified over 45 genomic risk loci for nsCL/P (Beaty et al., 2010, 2013; Birnbaum et al., 2009; Leslie et al., 2016, 2017; Ludwig et al., 2012, 2016, 2017; Mangold et al., 2010; Moreno et al., 2009; Mostowska et al., 2018; Mukhopadhyay et al., 2020, 2022; Rahimov et al., 2008; Sun et al., 2015; Welzenbach et al., 2021; Y. Yu et al., 2017). A large proportion of these discoveries relate to common variants identified by genome-wide association studies (GWAS). These studies determine the frequencies of genetic variants in individuals with a particular disease/trait (i.e. nsCL/P) in comparison to a control group (Ludwig et al., 2019). The statistical evaluation of GWAS provides a p-value for each variant, which describes the probability that the particular variant is associated with the disease/trait under investigation, and an effect size, which measures the strength of this association. However, most of the genetic associations for

nsCL/P identified by GWAS map to non-coding regions of the genome, as is typical for complex phenotypes (Thieme and Ludwig, 2017). This complicates the biological interpretation of these findings due to potential spatiotemporal effects of the associated variants on the expression of target genes (Maurano et al., 2012).

For this reason, follow-up studies are usually required to identify candidate genes that may be affected by regulatory effects of the identified non-coding variants, thereby providing more insight into the underlying biological mechanisms. Although challenging, different studies have determined candidate genes at some of the nsCL/P risk loci (Satokata and Maas, 1994; Thieme and Ludwig, 2017; Welzenbach et al., 2021). In addition to alternative strategies, one approach for the prioritization of candidate genes for GWAS risk loci is to utilize data on topologically associating domains (TADs). TADs are regions of the genome that physically interact with each other at a higher frequency than with regions located outside of a given TAD, which can potentially impact gene regulation of genes within the TADs (Pombo and Dillon, 2015). Consequently, it is likely that target genes of non-coding GWAS risk variants are located within the same TADs as the risk variants. Therefore, GWAS summary statistics can be combined with data on TADs to prioritize candidate genes based on non-coding risk variants. For nsCL/P, this approach was implemented by Welzenbach *et al*. 2021 by combining a GWAS meta-analysis with TAD data from human embryonic stem cells.

However, in order to interpret the biological relevance of the risk loci and to further elucidate the underlying mechanisms of nsCL/P, developmental cell types affected by the genetic variants need to be identified, as gene regulation can be cell type or tissue specific (Maurano et al., 2012). In human embryonic development, the development of the face takes place between the fourth and tenth week of gestation (Dixon et al., 2011). During this time period, cells from the neural crest (NC) undergo extensive migration and differentiation processes that lay the foundation of the structures that constitute the face (Cordero et al., 2011). The NC is a transient population of cells located at the dorsal part of the neural tube that possesses the ability to give rise to diverse cell types and tissues (Dooley et al., 2019). It is divided into four segments: cranial neural crest cells (CNCCs), cardiac, vagal and trunk neural crest cells. The CNCCs primarily give rise to the connective, skeletal, cartilage, bone and nerve tissue of the developing head (Roth et al.,

2021; Rothstein et al., 2018). Early in craniofacial development, CNCCs migrate to paired bulges on either side of the developing head, the pharyngeal arches (Graham et al., 2005). These develop the facial prominences that grow and fuse together in the facial midline, shaping the upper and lower jaws, chin, lips, philtrum, nose and palate (Danescu et al., 2015; Roth et al., 2021; Toro-Tobon et al., 2023). Despite continuous research, the finely orchestrated molecular processes underlying craniofacial development, as well as the developmental cell types relevant for nsCL/P, are still not fully understood.

Due to the spatiotemporal effects of gene regulation, gene expression patterns of nsCL/P candidate genes from relevant tissues can be used to identify potential cell types involved in nsCL/P. In the past, gene expression has been primarily examined using bulk RNA sequencing techniques, measuring the average gene expression in a sample, but missing potential gene expression heterogeneity between cell types within a tissue sample (Liao et al., 2022). However, the technological advancements of the last decade now allow the investigation of the transcriptomes of individual cells using various single-cell RNA sequencing (scRNA-seq) techniques (Sreenivasan et al., 2022). This innovation makes it possible to detect gene expression heterogeneity and discover previously unknown cell types and subpopulations, which has been demonstrated in the context of OFC by studies identifying transcriptional heterogeneity in the palate mesenchyme (Ozekin et al., 2023) and distinct cell populations at the fusion sites of the facial prominences in mice (Li et al., 2019).

The present thesis aimed to expand the interpretation of the genetic risk variants for nsCL/P by identifying potentially affected developmental cell types based on candidate gene expression patterns in high-resolution gene expression data. For this purpose, the expression patterns of candidate genes were examined in two independent murine scRNA-seq data sets from relevant tissues and time points (Cao et al., 2019; Li et al., 2019). First, an analysis workflow was implemented to re-analyze the published scRNA-seq data and to examine the expression patterns of GWAS candidate genes identified in Welzenbach *et al*. 2021 (Siewert et al., 2023). This analysis workflow was then also applied to study the expression patterns of candidate genes identified in whole-genome sequencing data (Zieger et al., 2023). While craniofacial development is quite conserved between humans and mice in general, there are also considerable differences between

the species (K. Yu et al., 2017). Therefore, to specifically identify human developmental cell types, an scRNA-seq data set from the heads of human embryos (Xu et al., 2023) was used for a more systematic analysis of nsCL/P candidate gene expression and cell type identification (Siewert et al., 2024).

REFERENCES

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., et al. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* 42, 525–529

Beaty, T. H., Taub, M. A., Scott, A. F., Murray, J. C., Marazita, M. L., Schwender, H., et al. (2013). Confirming genes influencing risk to cleft lip with/without cleft palate in a case-parent trio study. *Hum Genet* 132, 771–781

Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., et al. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41, 473–477

Blanton, S. H., Cortez, A., Stal, S., Mulliken, J. B., Finnell, R. H., and Hecht, J. T. (2005). Variation in IRF6 contributes to nonsyndromic cleft lip and palate. *Am J Med Genet* 137 A, 259–262

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502

Christensen, K., Juel, K., Herskind, A. M., and Murray, J. C. (2004). Long term follow up study of survival associated with cleft lip and palate at birth. *Br Med J* 328, 1405–1406

Cordero, D. R., Brugmann, S., Chu, Y., Bajpai, R., Jame, M., and Helms, J. A. (2011). Cranial neural crest cells on the move: Their roles in craniofacial development. *Am J Med Genet A* 155, 270–279

Danescu, A., Mattson, M., Dool, C., Diewert, V. M., and Richman, J. M. (2015). Analysis of human soft palate morphogenesis supports regional regulation of palatal fusion. *J Anat* 227, 474–486

Dixon, M. J., Marazita, M. L., Beaty, T. H., and Murray, J. C. (2011). Cleft lip and palate: Understanding genetic and environmental influences. *Nat Rev Genet* 12, 167–178

Dooley, C. M., Wali, N., Sealy, I. M., White, R. J., Stemple, D. L., Collins, J. E., et al. (2019). *The gene regulatory basis of genetic compensation during neural crest induction*

Dunkhase, E., Ludwig, K. U., Knapp, M., Skibola, C. F., Figueiredo, J. C., Hosking, F. J., et al. (2016). Nonsyndromic cleft lip with or without cleft palate and cancer: Evaluation of a possible common genetic background through the analysis of GWAS data. *Genom Data* 10, 22–29

Gajdos, V., Bahuau, M., Robert-Gnansia, E., Francannet, C., Cordier, S., and Bonaïti-Pellié, C. (2004). Genetics of nonsyndromic cleft lip with or without cleft palate: Is there a Mendelian sub-entity? *Ann Genet* 47, 29–39

Graham, A., Okabe, M., and Quinlan, R. (2005). The role of the endoderm in the development and evolution of the pharyngeal arches. *J Anat* 207, 479–487

Grosen, D., Bille, C., Petersen, I., Skytthe, A., Hjelmborg, J. V. B., Pedersen, J. K., et al. (2011). Risk of oral clefts in twins. *Epidemiology* 22, 313–319

Jezewski, P. A., Vieira, A. R., Nishimura, C., Ludwig, B., Johnson, M., O'Brien, S. E., et al. (2003). Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. *J Med Genet* 40, 399–407

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., et al. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32, 285–289

Kousa, Y. A., and Schutte, B. C. (2016). Toward an orofacial gene regulatory network. *Developmental Dynamics* 245, 220–232

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Butali, A., Buxó, C. J., Castilla, E. E., et al. (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Hum Genet* 136, 275–286

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Feingold, E., Wehby, G., Laurie, C. A., et al. (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p 24.2, 17q23 and 19q13. *Hum Mol Genet* 25, 2862–2872

Li, H., Jones, K. L., Hooper, J. E., and Williams, T. (2019). The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. *Development (Cambridge)* 146

Liao, J., Qian, J., Fang, Y., Chen, Z., Zhuang, X., Zhang, N., et al. (2022). De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution. *Nat Commun* 13, 1–19

Ludwig, K. U., Ahmed, S. T., Böhmer, A. C., Sangani, N. B., Varghese, S., Klamt, J., et al. (2016). Meta-analysis Reveals Genome-Wide Significance at 15q13 for Nonsyndromic Clefting of Both the Lip and the Palate, and Functional Analyses Implicate GREM1 As a Plausible Causative Gene. *PLoS Genet* 12, 1–21

Ludwig, K. U., Böhmer, A. C., Bowes, J., Nikolić, M., Ishorst, N., Wyatt, N., et al. (2017). Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. *Hum Mol Genet* 26, 829–842

Ludwig, K. U., Degenhardt, F., and Nöthen, M. M. (2019). The role of rare variants in common diseases. *Medizinische Genetik* 31, 212–221

Ludwig, K. U., Mangold, E., Herms, S., Nowak, S., Paul, A., Becker, J., et al. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. 44, 968–971

Mangold, E., Böhmer, A. C., Ishorst, N., Hoebel, A. K., Gültepe, P., Schuenke, H., et al. (2016). Sequencing the GRHL3 Coding Region Reveals Rare Truncating Mutations and a Common Susceptibility Variant for Nonsyndromic Cleft Palate. *Am J Hum Genet* 98, 755–762

Mangold, E., Kreiß, M., and Nöthen, M. M. (2017). Syndromale und nichtsyndromale orofaziale Spalten. *Medizinische Genetik* 29, 397–412

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., et al. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42, 24–26

Mangold, E., Ludwig, K. U., and Nöthen, M. M. (2011). Breakthroughs in the genetics of orofacial clefting. *Trends Mol Med* 17, 725–733

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Available at: https://www.science.org

Moreno, L. M., Mansilla, M. A., Bullard, S. A., Cooper, M. E., Busch, T. D., Machida, J., et al. (2009). FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Hum Mol Genet* 18, 4879–4896

Mostowska, A., Gaczkowska, A., Żukowski, K., Ludwig, K. U., Hozyasz, K. K., Wójcicki, P., et al. (2018). Common variants in DLG1 locus are associated with non-syndromic cleft lip with or without cleft palate. *Clin Genet* 93, 784–793

Mukhopadhyay, N., Bishop, M., Mortillo, M., Chopra, P., Hetmanski, J. B., Taub, M. A., et al. (2020). 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

Mukhopadhyay, N., Feingold, E., Moreno-Uribe, L., Wehby, G., Valencia-Ramirez, L. C., Restrepo Muñeton, C. P., et al. (2022). Genome-wide association study of multiethnic nonsyndromic orofacial cleft families identifies novel loci specific to family and phenotypic subtypes. *Genet Epidemiol* 46, 182–198

Ozekin, Y. H., O'Rourke, R., and Bates, E. A. (2023). Single cell sequencing of the mouse anterior palate reveals mesenchymal heterogeneity. *Developmental Dynamics* 252, 713–727

Pengelly, R. J., Arias, L., Martinez, J., Upstill-Goddard, R., Seaby, E. G., Gibson, J., et al. (2016). Deleterious coding variants in multi-case families with non-syndromic cleft lip and/or palate phenotypes. *Sci Rep* 6, 1–8

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: Players and mechanisms. *Nat Rev Mol Cell Biol* 16, 245–257

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., et al. (2008). Disruption of an AP-2α binding site in an IRF6 enhancer is strongly associated with cleft lip. *Nat Genet* 40, 1341–1347

Roth, D. M., Bayona, F., Baddam, P., and Graf, D. (2021). Craniofacial Development: Neural Crest in Molecular Embryology. *Head Neck Pathol* 15, 1–15

Rothstein, M., Bhattacharya, D., and Simoes-Costa, M. (2018). The molecular basis of neural crest axial identity. *Dev Biol* 444, S170–S180

Satokata, I., and Maas, R. (1994). Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. *Nat Genet* 6.

Siewert, A., Hoeland, S., Mangold, E., and Ludwig, K. U. (2024). Combining genetic and single-cell expression data reveals cell types and novel candidate genes for orofacial clefting. *Sci Rep* 14 26492

Siewert, A., Reiz, B., Krug, C., Heggemann, J., Mangold, E., Dickten, H., et al. (2023). Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data. *Front Cell Dev Biol* 11, 1–11

Sreenivasan, V. K. A., Balachandran, S., and Spielmann, M. (2022). The role of single-cell genomics in human genetics. *J Med Genet* 59, 827–839

Sun, Y., Huang, Y., Yin, A., Pan, Y., Wang, Y., Wang, C., et al. (2015). Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat Commun* 6

Thieme, F., and Ludwig, K. U. (2017). The Role of Noncoding Genetic Variation in Isolated Orofacial Clefts. *J Dent Res* 96, 1238–1247

Toro-Tobon, S., Paredes-Gutierrez, J., Mantilla-Rivas, E., Ahmad, L., and Rogers, G. F. (2023). Pharyngeal Arches, Chapter 1: Normal Development and Derivatives. *Journal of Craniofacial Surgery* 34, 2237–2241

van der Woude, A. (1954). Fistula labii inferioris congenita and its association with cleft lip and palate. *Am J Hum Genet* 6, 244–256.

Welzenbach, J., Hammond, N. L., Nikolić, M., Thieme, F., Ishorst, N., Leslie, E. J., et al. (2021). Integrative approaches generate insights into the architecture of non-syndromic cleft lip with or without cleft palate. *Human Genetics and Genomics Advances* 2, 1–14

Xu, Y., Zhang, T., Zhou, Q., Hu, M., Qi, Y., Xue, Y., et al. (2023). A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nat Cell Biol*, 1–12

Yu, K., Deng, M., Naluai-Cecchini, T., Glass, I. A., and Cox, T. C. (2017a). Differences in oral structure and tissue interactions during mouse vs. human palatogenesis: Implications for the translation of findings from mice. *Front Physiol* 8, 1–12

Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., et al. (2017b). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat Commun* 8, 1–11

Zieger, H. K., Weinhold, L., Schmidt, A., Holtgrewe, M., Juranek, S. A., Siewert, A., et al. (2023). Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of de novo mutations. *Human Genetics and Genomics Advances* 4, 100166

4 Publications

4.1 Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data

Anna Siewert, Benedikt Reiz, Carina Krug, Julia Heggemann, Elisabeth Mangold, Henning Dickten and Kerstin U. Ludwig

Frontiers in Cell and Developmental Biology 11:1091666 (2023)

https://doi.org/10.3389/fcell.2023.1091666

frontiers | Frontiers in Cell and Developmental Biology

TYPE Original Research
PUBLISHED 24 April 2023
DOI 10.3389/fcell.2023.1091666

OPEN ACCESS

EDITED BY
Walid D. Fakhouri,
University of Texas Health Science Center at Houston, United States

REVIEWED BY
Hong Li,
University of Colorado Anschutz Medical Campus, United States
Ariadne Letra,
University of Pittsburgh, United States

*CORRESPONDENCE
Kerstin U. Ludwig,
kerstin.ludwig@uni-bonn.de

RECEIVED 07 November 2022
ACCEPTED 03 April 2023
PUBLISHED 24 April 2023

CITATION
Siewert A, Reiz B, Krug C, Heggemann J, Mangold E, Dickten H and Ludwig KU (2023), Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data.
Front. Cell Dev. Biol. 11:1091666.
doi: 10.3389/fcell.2023.1091666

Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data

Anna Siewert[1], Benedikt Reiz[2], Carina Krug[1], Julia Heggemann[1], Elisabeth Mangold[1], Henning Dickten[2] and Kerstin U. Ludwig[1]*

[1]Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn, Germany, [2]FASTGenomics, Comma Soft AG, Bonn, Germany

Introduction: Cleft lip ± cleft palate (CL/P) is one of the most common birth defects. Although research has identified multiple genetic risk loci for different types of CL/P (i.e., syndromic or non-syndromic forms), determining the respective causal genes and understanding the relevant functional networks remain challenging. The recent introduction of single-cell RNA sequencing (scRNA-seq) has provided novel opportunities to study gene expression patterns at cellular resolution. The aims of our study were to: (i) aggregate available scRNA-seq data from embryonic mice and provide this as a resource for the craniofacial community; and (ii) demonstrate the value of these data in terms of the investigation of the gene expression patterns of CL/P candidate genes.

Methods and Results: First, two published scRNA-seq data sets from embryonic mice were re-processed, i.e., data representing the murine time period of craniofacial development: (i) facial data from embryonic day (E) E11.5; and (ii) whole embryo data from E9.5–E13.5 from the Mouse Organogenesis Cell Atlas (MOCA). Marker gene expression analyses demonstrated that at E11.5, the facial data were a high-resolution representation of the MOCA data. Using CL/P candidate gene lists, distinct groups of genes with specific expression patterns were identified. Among others we identified that a co-expression network including *Irf6, Grhl3* and *Tfap2a* in the periderm, while it was limited to *Irf6* and *Tfap2a* in palatal epithelia, cells of the ectodermal surface, and basal cells at the fusion zone. The analyses also demonstrated that additional CL/P candidate genes (e.g., *Tpm1, Arid3b, Ctnnd1, and Wnt3*) were exclusively expressed in *Irf6*+ facial epithelial cells (i.e., as opposed to *Irf6*- epithelial cells). The MOCA data set was finally used to investigate differences in expression profiles for candidate genes underlying different types of CL/P. These analyses showed that syndromic CL/P genes (syCL/P) were expressed in significantly more cell types than non-syndromic CL/P candidate genes (nsCL/P).

Discussion: The present study illustrates how scRNA-seq data can empower research on craniofacial development and disease.

KEYWORDS

cleft lip with or without cleft palate, single-cell RNA sequencing (scRNA-seq), IRF6, craniofacial development, expression pattern, single-cell transcriptomics

1 Introduction

Molecular malfunctions during craniofacial development can lead to cleft lip ± cleft palate (CL/P). CL/P represents one of the most common of all birth defects, with a global prevalence of 1 in 700 live births (Mangold et al., 2011). Importantly, CL/P can present either as an isolated, non-syndromic phenotype (nsCL/P), or within the context of more complex malformation syndromes (syCL/P), in which additional features indicative of a developmental defect are observed. Although CL/P can be caused by deleterious mutations in single high penetrance genes (Cox et al., 2018; Bishop et al., 2020), a considerable fraction of its genetic architecture is attributable to common risk variants. Research suggests that environmental factors also contribute to CL/P, as part of its multifactorial etiology (Murray, 2002).

For nsCL/P, genome-wide association studies (GWAS) have identified multiple risk loci, and positional analyses of these loci have revealed promising candidate genes. For most of these genes, however, few data are available concerning the mechanism through which they affect the underlying functional processes of craniofacial development. One of the few exceptions to this is the *IRF6-GRHL3-TFAP2A* network, which has been shown to underlie diverse types of orofacial clefting, including CL/P and cleft palate only (Kousa et al., 2019). In addition to challenges associated with attributing causality to individual variants and genes, this lack of knowledge is also explained by the limited access to molecular data from relevant time points in humans, due to technical and ethical limitations.

Recently, single-cell RNA sequencing (scRNA-seq) has been performed on tissue from embryonic mice, generating systematic transcriptomic data sets at cellular resolution. This offers new avenues for the study of the tissue-specific expression of genes that underlie developmental phenotypes, including CL/P. Two resources of particular value in terms of CL/P are the Mouse Organogenesis Cell Atlas (MOCA; Cao et al., 2019), and facial data from embryonic mice that were reported in 2019 (Li et al., 2019). While MOCA encompasses the developmental time frame embryonic day (E) 9.5–13.5, the data from Li et al. provide a deeper insight into the transcriptome of facial structures at E11.5. Two important challenges associated with the use of scRNA-seq data are data accessibility and comparability, particularly when data are generated in different labs. The data of MOCA and Li et al. vary in terms of the level of processing, output types, and usability for the research community.

The aims of the present study were to (i) aggregate these scRNA-seq data from embryonic mice and provide this as a resource for the craniofacial community; and (ii) demonstrate the value of these data in terms of the investigation of the gene expression patterns of CL/P candidate genes. First, both of the selected data sets were re-analyzed using a joint computational pipeline. Second, different CL/P candidate gene sets were used to illustrate the potential of scRNA-seq data for deciphering the CL/P etiology. In particular, the expression patterns of CL/P candidate genes were assessed across the time period of craniofacial development, with the aim of placing them in their cell type-specific context. We specifically analyzed epithelial and mesenchymal cell types, which have been previously shown to be involved in CL/P (Ji et al., 2020). As an application example, we investigated co-expression of members of the *Irf6-Grhl3-Tfap2a* genetic pathway in epithelial cell sub-types and identified further genes with a potential *Irf6* interaction in these cells. Finally, potential expression differences in candidate genes for syCL/P and nsCL/P were investigated in order to test the hypothesis that during embryonic development, syCL/P candidate genes are expressed in more tissues than is the case for candidate genes for nsCL/P.

2 Materials and methods

2.1 Data sources

Two sets of single cell data on murine embryonic development were downloaded and analyzed using the same computational pipeline, which is described in detail in "Data analysis." The first data set comprised single-cell gene expression data from 7,893 single cells from the lambdoidal junction, which were extracted from 4-5 mouse embryos at E11.5 (Li et al., 2019). The corresponding gene-count matrix was downloaded from the Gene Expression Omnibus (RRID:SCR_005012; accession number: GSM3867275). The data set was then re-analyzed using our in-house pipeline. The latter included stricter filtering parameters (see below), thus reducing the number of single cells used for analysis (7,249 cells in total) compared to the original study. The final facial data set included 25 cell clusters.

The second data set was MOCA, which was generated from whole embryonic mice (Cao et al., 2019). The MOCA data set comprises the expression data of 2,058,652 single cells, as obtained from 61 mouse embryos from developmental stages E9.5–E13.5. Post-filtering, the original data set contained data on 1,331,985 cells and 38 major cell types (Supplementary Table S1). The gene-count matrix containing these 1,331,985 pre-filtered, high-quality cells was downloaded from the MOCA Website, and stored and analyzed using FASTGenomics (Scholz et al., 2018; RRID:SCR_022898). In contrast to the original publication, the gene count matrix was split into five data sets in accordance with embryonic day in order to create a developmental time frame of gene expression: 112,269 cells (E9.5); 258,104 cells (E10.5); 449,614 cells (E11.5); 270,197 cells (E12.5); and 241,800 cells (E13.5).

2.2 Data analysis

2.2.1 General processing

Each of the data sets was processed using the R package Seurat v4 (Hao et al., 2021; RRID:SCR_016341). To normalize the count matrices, Log normalization (normalization.method) was applied with a Seurat default scale factor of 10,000 (scale.factor). For the selection of highly variable genes, the "vst" selection method (selection.method) was chosen, using 2,500 as the number of features (nfeatures). Scaling was performed in block sizes of 500 (block.size). For linear dimension reduction, a principal component (PC) analysis was performed. To cluster the cells, a two-step approach was used. First, for each cell, the K-nearest neighbors were calculated using the *FindNeighbors* function of Seurat, based on the first 25 PC dimensions (dims). Second, the Louvain algorithm was applied as a modularity optimization technique with a resolution of 0.5 for MOCA data and 1.1 for facial data (resolution) using the *FindClusters* function. To identify differentially expressed genes (hereafter

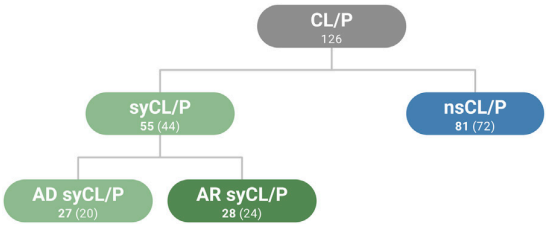


FIGURE 1
Summary of gene lists used in the present study. Genes that overlapped between categories are included in the numbers of genes indicated in bold (n = 10). Numbers in parentheses correspond to the number of unique genes in the respective category, without overlapping genes. CL/P (cleft lip with or without cleft palate), ns (non-sndromic), sy (syndromic), AD (autosomal dominant), AR (autosomal recessive).

referred to as 'marker genes') for each cluster, the Wilcoxon Rank Sum test was used (test.use). Marker genes were obtained by comparing the expression levels of individual genes against all other clusters, and only positive markers were used. Additional parameters were a minimum fraction of 0.25 of cells expressing the tested gene in either of the populations (min.pct), and a threshold of a 0.25-fold change between the tested clusters (logfc.threshold). The uniform manifold approximation and projection algorithm (UMAP) was used as a non-linear dimension reduction method, whereby the first 25 PCs were applied as dimensions (dims).

2.2.2 Study-specific filtering

For the facial data set, additional steps were performed pre-normalization. These included the filtering-out of potential doublets by excluding cells with >7,500 unique features (nFeature_RNA), and cells with >80,000 detected RNA molecules (nCount_RNA). To exclude cells that were previously lysed or apoptotic, cells with the presence of the following features were excluded from the data set: (i) a percentage of >5% of unique molecular identifiers reflecting mitochondrial genes (percent.mt); and/or (ii) < 2,300 unique features (nFeature_RNA). After filtering, our data set comprised 7,249 cells. To benchmark the present pipeline, cell type annotation was performed by comparing the marker genes of each cluster with the marker genes described in the original publication.

For the pre-filtered, high-quality cells of the MOCA data, no additional filtering was required. Final cell type annotation was performed using the published marker genes of Cao et al. and the R package scCATCH (Shao et al., 2020). For the latter, *species* was set to "Mouse"; *match_CellMatch* was set to "TRUE"; and the tissues selected to be matched to "CellMatch" were "Brain," "Fetal brain" and "Embryo". Further parameters were kept at default values.

2.3 Curation of CL/P candidate gene lists

A literature search was performed to generate lists of genes associated with non-syndromic and syndromic forms of CL/P. The nsCL/P gene list was generated based on a recent meta-analysis of nsCL/P GWAS (Welzenbach et al., 2021). Welzenbach et al. performed a gene-based analysis for genes located at established GWAS risk loci, which identified a set of 81 genes with an enrichment of common variants. These 81 genes were used in the present study. The syCL/P *gene list* was generated using information from a recently published study (Bishop et al., 2020), which had involved a systematic review of orofacial clefting syndromes and their associated genes. For the purposes of the present study, the list of syndromes generated by Bishop et al. was reduced using OMIM (RRID:SCR_006437) in order to: (i) include only those syndromes whose phenotype includes CL/P, with the exclusion of other orofacial clefting phenotypes; and (ii) generate subsets of genes with autosomal dominant (AD) or autosomal recessive (AR) contributions. An overview of the gene categories is provided in Figure 1. Genes that overlapped between the syndromic and non-syndromic categories were included in an 'overlapping genes' list. Use of this list was restricted to the comparison of expression data between syCL/P and nsCL/P. The final numbers of unique genes for these analyses were 126 genes for CL/P overall, of which 72 genes were for nsCL/P, and 44 genes were for syCL/P (20 AD genes and 24 AR genes). Ten genes overlapped both categories.

To evaluate whether the findings for CL/P are generalizable to other birth defects, gene lists were also generated for congenital heart disease (CHD). A recent publication (Nees and Chung, 2020) listed 18 genes for non-syndromic CHD (nsCHD) and 56 genes for syndromic CHD (40 AD, 16 AR). Three genes overlapped both categories. However, this group was not analyzed in the present study due to the low number of genes. All gene lists are provided in Supplementary Table S2.

2.4 Creating Irf6+ and Irf6- epithelial cell sub-clusters

Based on its well-established role in both syCL/P and nsCL/P (Woude, 1954; Birnbaum et al., 2009), analyses were performed to investigate the role of *Irf6* in epithelial cells. To create *Irf6*+ and *Irf6*- epithelial sub-clusters, epithelial cell clusters in the facial data set (i.e., palatal epithelium, olfactory epithelium, ectodermal surface, ectodermal surface (*Robo2*+), periderm, and basal cells at the fusion zone) were divided into subsets according to *Irf6* expression. Previous research has shown that *Irf6*, *Grhl3,* and *Tfap2a* are part of a genetic network in which *Irf6* influences the gene expression of *Grhl3* and *Tfap2a* (Kousa et al., 2019). In order to examine if these genes are among the marker genes of the *Irf6*+ sub-clusters and to identify possible additional genes that are influenced by *Irf6*, we determined marker genes for these sub-clusters. For this purpose, the expression profiles of each sub-cluster were compared against all other cell clusters in the data set, using the parameters applied in the initial data analysis (see Data analysis; Supplementary Table S3).

2.5 Analysis of differences in gene expression between nsCL/P and syCL/P

The analysis of nsCL/P and syCL/P gene lists was performed in the whole embryo MOCA data sets. Two parameters were used in these comparisons: (i) the percentage of all cell types in which the respective genes were expressed; and (ii) the average expression level. For analysis (i), a cell type was considered to express a certain

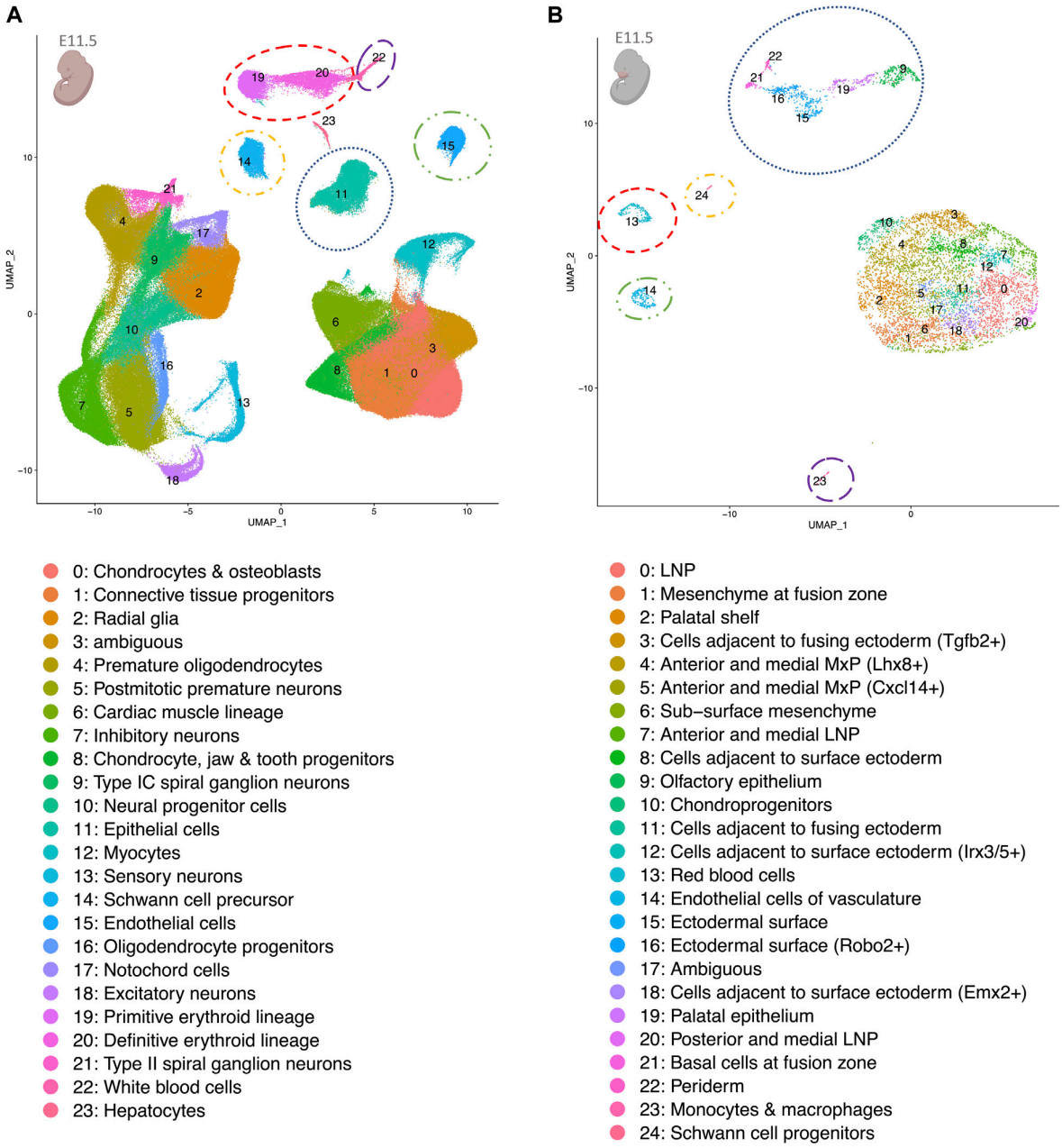


FIGURE 2
UMAP plots of re-analyzed scRNA-seq whole embryo data at E11.5 **(A)** and facial data at E11.5 **(B)**. Despite differing read depths in the two data sets, shared cell clusters corresponding to matched cell types are observed. These are encircled in the same color in both panels. The pink colors of the embryo graphics correspond to the tissues that are included in the data set. Lateral nasal process (LNP), maxillary prominence (MxP).

gene if the gene was expressed in at least 10% of cells. Percentages were determined for each gene in the respective list. The distributions were statistically compared using the Welch *t*-test. For analysis (ii), the average expression levels per cell type were extracted for each gene using the *AverageExpression* function from Seurat v4. The mean of these expression levels was then calculated per gene. A statistical comparison of the mean expression levels between both gene lists was performed using the Welch *t*-test.

3 Results

3.1 Facial-specific and whole embryo scRNA-seq data provide complementary insights into craniofacial development

Figure 2 shows the results generated by the UMAP algorithm for both the facial data (panel B) and the MOCA data (panel A, E11.5,

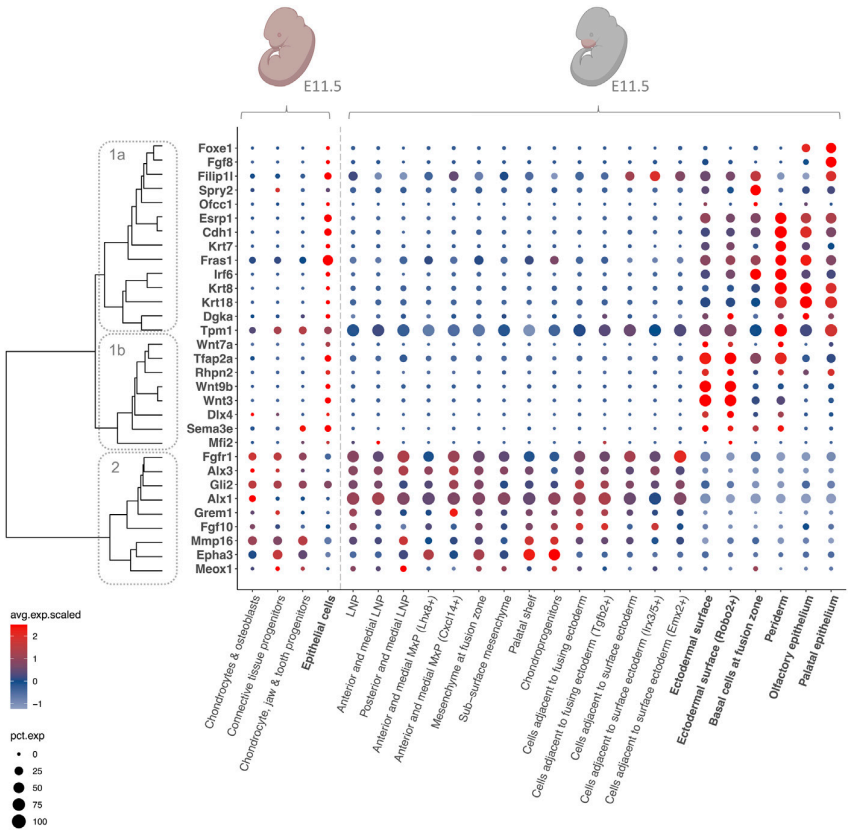


FIGURE 3
Expression patterns of CL/P candidate genes at E11.5. Dotplot of gene expression for selected CL/P candidate genes at E11.5 in selected cell types in the MOCA and the facial data. The color of the dots corresponds to the average scaled expression level. The size of the dots corresponds to the percentage of cells that express the gene in the respective cell type. Epithelial cell types are indicated in bold, and mesenchymal-like cell types are indicated in non-bold. Dendrogram cluster 1a = genes expressed predominantly in periderm, basal cells at fusion zone, olfactory epithelium, and palatal epithelium; cluster 1b = genes predominantly expressed in ectodermal surface; cluster 2 = genes expressed predominantly in mesenchymal-like cell types. The pink colors of the embryo graphics correspond to the tissues that are included in the data set. Lateral nasal process (LNP), maxillary prominence (MxP).

all other time points in Supplementary Figures S1A–D). The 25 cell types observed in the facial data were grouped into two main cell type clusters: (i) epithelial cells comprising periderm, basal cells at fusion zone, ectodermal surface, ectodermal surface (*Robo2*+), olfactory epithelium, and palatal epithelium; and (ii) more diverse cell types, which share a mesenchymal state, as based on the analysis of mesenchymal cell markers (Supplementary Figure S1E). Smaller cell clusters included endothelial cells and Schwann cells (Figure 2B). In the MOCA data for E11.5, a total of 24 cell types were identified, including a distinct epithelial cluster. To determine whether this at least partially represents the epithelial clusters in the facial data, the epithelial cells were sub-clustered. Three of these sub-clusters express marker genes for periderm (sub-cluster 6), basal cells at the fusion zone (sub-cluster 7) and ectodermal surface (sub-clusters 8 & 9) (Supplementary Figure S1G; marker genes of the sub-clusters in Supplementary Table S3). Additional cell clusters in MOCA comprised specific cell types, such as hepatocytes, which are not represented in the facial data, as well as overlapping cell types where expected, e.g., endothelial cells, Schwann cells, and red and white blood cell types (Figure 2A, B colored circles).

3.2 A subset of CL/P candidate genes show convergent expression patterns

Investigation of the expression patterns of CL/P candidate genes in the scRNA-seq data sets showed that while the facial data set allowed an in-depth investigation of craniofacial structures at E11.5, the MOCA data set enabled a time course analysis over the time span of craniofacial development. Of the 126 CL/P candidate genes, all were expressed in the MOCA data sets from E9.5 - E13.5, although they varied in terms of overall expression levels and the cell types in which they were expressed. In the MOCA data, many CL/P candidate genes showed ubiquitous expression at E9.5, which became more specific at E10.5. Among the 126 CL/P candidate genes, 31 were specifically expressed in cell types of relevance to craniofacial development (i.e., epithelial cells, chondrocytes and osteoblasts, connective tissue progenitors, chondrocyte and jaw and tooth progenitors). Here, “specific expression” refers to either: (i) expression in at least one of these cell types; or (ii) expression in additional cell types, but with the highest expression levels being observed in at least one of the cell types

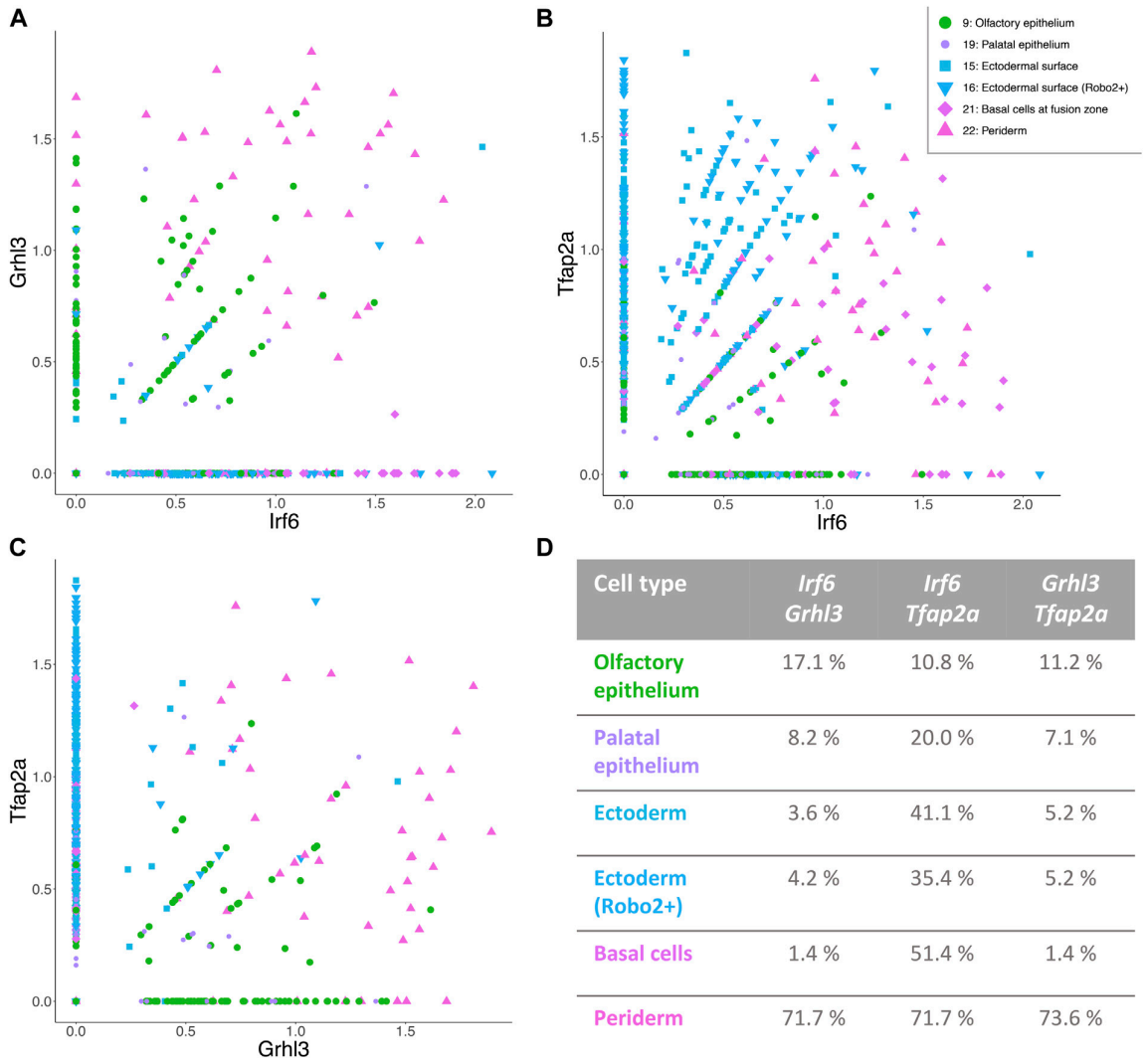
TABLE 1 CL/P candidate genes with specific expression in *Irf6*+ facial epithelial cells.[1] adjusted *p*-value (based on Bonferroni correction using all genes in the data set);[2] average log_2 fold change in the average expression between the two tested groups (second test group: all other cell types; positive values indicate that the gene is more highly expressed in the respective cell type compared to all other cell types). NsCLO (non-syndromic cleft lip only).

Cell type	Gene	P-val. adj.[1]	Log2FC[2]	Cleft association in humans
Periderm	*Tpm1*	1.9E-09	1.08	nsCL/P GWAS (Ludwig et al., 2012)
	Pik3r1	6.1E-10	0.75	nsCL/P GWAS (Leslie et al., 2017)
	Tfap2a	4.9E-44	1.03	Branchio-oculo-facial syndrome (Milunsky et al., 2008), nsCL/P GWAS (Ludwig et al., 2012; Leslie et al., 2017)
	Wnt3	1.0E-10	0.25	Tetra-amelia syndrome 1 (Niemann et al., 2004)
	Ctnnd1	0.0002	0.42	Blepharocheilodontic syndrome 2 (Ghoumid et al., 2017)
	Fras1	8.0E-10	0.69	Fraser syndrome (Fraser, 1962)
Basal cells at fusion zone	*Spry2*	4.0E-37	1.29	nsCL/P GWAS (Ludwig et al., 2012)
Ectodermal surface	*Arid3b*	7.7E-13	0.3	nsCL/P GWAS (Leslie et al., 2017)
	Zfp36l2	0.008	0.26	nsCL/P (Lin-Shiao et al., 2019), nsCLO (Li et al., 2022)
Ectodermal surface (*Robo2*+)	*Tpm1*	0.04	0.3	nsCL/P GWAS (Ludwig et al., 2012)
Palatal epithelium	*Cyb561*	1.6E-36	0.3	nsCL/P GWAS (Leslie et al., 2016)
	Ptch1	0.0001	0.33	nsCL/P GWAS (Yu et al., 2017), CPO GWAS (Butali et al., 2019), Basal cell nervous syndrome (Evans et al., 1993; Kimonis et al., 1997; Kimonis et al., 2013)
	Tfap2a	1.9E-09	0.27	Branchio-oculo-facial syndrome (Milunsky et al., 2008), nsCL/P GWAS (Ludwig et al., 2012; Leslie et al., 2017)
	Fras1	2.0E-08	0.33	Fraser syndrome (Fraser, 1962)
	Ripk4	5.3E-32	0.3	Popliteal pterygium syndrome, Bartsocas-Papas type 1 (Bartsocas and Papas, 1972)
Olfactory epithelium	*Arid3b*	2.5E-11	0.25	nsCL/P GWAS (Leslie et al., 2017)
	Cyb561	5.0E-55	0.29	nsCL/P GWAS (Leslie et al., 2016)
	Xpa	0.0003	0.26	nsCL/P GWAS (Welzenbach et al., 2021)

of relevance to craniofacial development. Comparison of the expression patterns of these 31 genes in the MOCA and the facial data (Figure 3) revealed that they clustered into two main groups: While 22 genes were specifically expressed in epithelial cell types (Figure 3 dendrogram cluster 1), nine genes were expressed in mesenchymal-like cell types (Figure 3 dendrogram cluster 2). Interestingly, the analyses showed that the first group (i.e., genes expressed predominantly in epithelial cell types) can be further subdivided into genes that have their highest expression levels in the ectodermal surface (Figure 3 dendrogram cluster 1b), and genes that have their highest expression levels in periderm, basal cells at fusion zone, olfactory epithelium, and palatal epithelium (Figure 3 dendrogram cluster 1a). The expression patterns of the remaining 95 CL/P candidate genes at E11.5 are shown in Supplementary Figure S4.

3.3 CL/P may involve distinct subgroups of epithelial cells

Using our data set, we first focused on the well-established CL/P risk gene *IRF6*. In the present study, *Irf6* was predominantly expressed in epithelial cells in both the MOCA and the facial data sets, with particularly strong expression being observed in the periderm and basal cells at fusion zone in the facial data set. In MOCA, this expression was maintained throughout the developmental time period of the data set (Supplementary Figure S3). In the facial epithelial cells, considerable intra-cluster heterogeneity was observed. Cells expressing *Irf6* (denoted as *Irf6*+ cells) were observed in 58% of cells from the palatal epithelium (n = 71 out of 170 cells), 40% of cells from olfactory epithelium (105/258), 44% of cells from the ectodermal surface (85/192), 44% of cells from the ectodermal surface (*Robo2*+) (86/192), 70% of cells from the basal cells at fusion zone (49/70), and 77% of the periderm cells (41/53). The six epithelial cell clusters from the facial data set were each divided into subsets according to their expression of *Irf6*, and marker genes of the *Irf6*+ cells were identified (Supplementary Table S3). A set of genes that overlapped between the marker genes of the *Irf6*+ epithelial subsets and CL/P candidate genes was identified (Table 1), which included CL/P genes that were associated with: (i) syndromic forms (e.g., *Tfap2a* (Branchio-oculo-facial syndrome, Milunsky et al., 2008), *Ctnnd1* (Blepharocheilodontic syndrome 2, Ghoumid et al., 2017) and *Fras1* (Fraser syndrome 1, Fraser, 1962); and (ii) candidate genes from GWAS loci (e.g., *Tpm1* (Ludwig et al., 2012) and *Arid3b* (Leslie et al., 2017) (Table 1,



Cell type	*Irf6 Grhl3*	*Irf6 Tfap2a*	*Grhl3 Tfap2a*
Olfactory epithelium	17.1 %	10.8 %	11.2 %
Palatal epithelium	8.2 %	20.0 %	7.1 %
Ectoderm	3.6 %	41.1 %	5.2 %
Ectoderm (Robo2+)	4.2 %	35.4 %	5.2 %
Basal cells	1.4 %	51.4 %	1.4 %
Periderm	71.7 %	71.7 %	73.6 %

FIGURE 4
Distinct populations of epithelial cells with a possible involvement in CL/P. **(A–C)** *Irf6-Grhl3-Tfap2a* show partial co-expression in epithelial cell types of E11.5 facial data. The axes of the graphs represent the expression level. Legend for all three figures is positioned in panel **(B)**. **(D)** Table showing the percentage of cells with co-expression of the respective gene pair in all six epithelial cell clusters.

Supplementary Table S4). Interestingly, the gene Grainyhead-like 3 (*Grhl3*) was also observed among the marker genes of cells from the periderm and olfactory epithelium. As with *Irf6*, mutations in *Grhl3* cause Van der Woude syndrome. Here, however, most individuals present with a cleft palate only rather than CL/P (Mangold et al., 2016).

To elucidate the connection of *Irf6*, *Grhl3*, and *Tfap2a* in the six epithelial cell types at the transcriptomic level, the co-expression of these genes was analyzed (Figures 4A–D). Each of the *Irf6-Grhl3-Tfap2a* gene pairs showed partial co-expression, since an overlap in expression was observed in a subgroup of cells (indicated by percentage in Figure 4D). The co-expression network comprising all three genes was most abundant in the periderm, while it was reduced to only *Irf6* and *Tfap2a* in basal cells at fusion zone, the ectodermal surface clusters, and the palatal epithelium as well.

3.4 SyCL/P genes are expressed in more tissues compared to nsCL/P genes

To compare differences in the number of cell types between the gene lists for syCL/P and nsCLP, the analysis was restricted to the MOCA data set only, since syCL/P can affect tissues and organs outside of the craniofacial region and the MOCA data set contains more non-facial tissues. Across stages E10.5 to E11.5, the syCL/P genes were expressed in significantly more cell types than was the case for the nsCL/P genes (Figure 5A, E11.5). Comparison of the average gene expression levels of these gene sets showed that the syCL/P genes did not have significantly higher gene expression levels than the nsCL/P genes (Figure 5B E11.5). However, division of the syCL/P gene set into AD and AR genes revealed that the observed differences

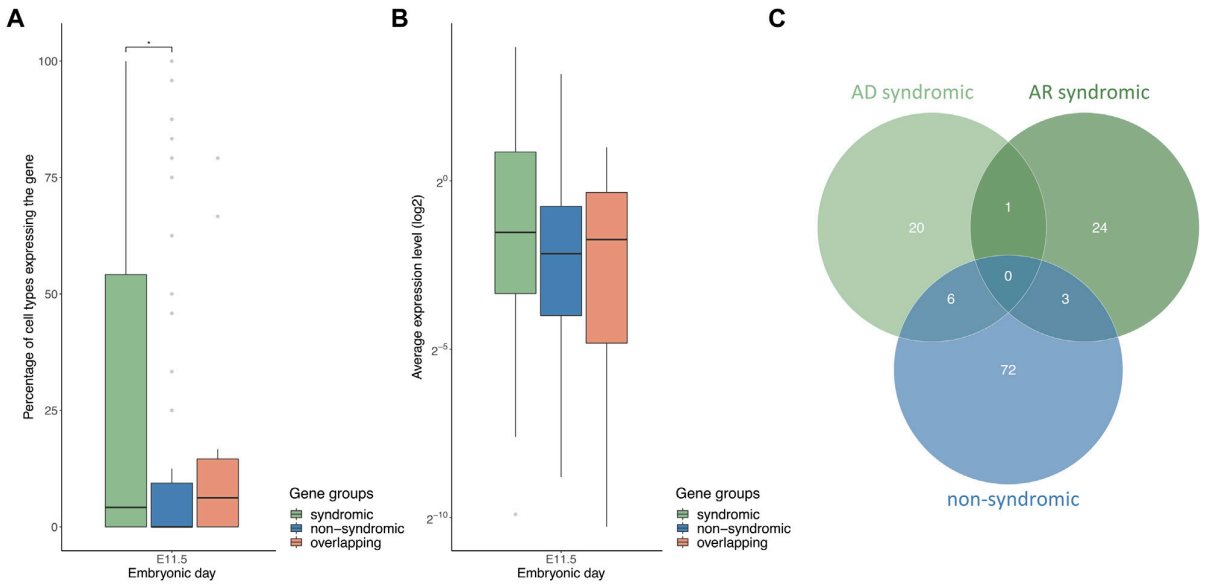


FIGURE 5
AD syCL/P genes are expressed in more cell types and have higher average expression levels compared with nsCL/P genes. **(A)** Boxplot of the percentages of cell types expressing the gene groups of syCL/P, nsCL/P, and overlapping genes at E11.5 **(B)** Boxplot of average log2 expression levels of the gene groups of syCL/P, nsCL/P, and overlapping genes at E11.5. **(C)** Venn diagram of non-syndromic, AR syndromic, and AD syndromic CL/P gene lists. (*$p < 0.05$). Data on the remaining time points are provided in Supplementary Figure S2.

in the percentage of expressing cell types between syCL/P and nsCL/P were mainly driven by the AD syCL/P genes. In comparison to the nsCL/P and AR syCL/P genes, the AD syCL/P genes: (i) were expressed in more cell types (Supplementary Figure S2A); and (ii) showed higher average expression levels (Supplementary Figure S2B).

4 Discussion

The present study leveraged two scRNA-seq data sets to generate insights into craniofacial development and diseases, specifically CL/P. Our reasons for selecting these data sets were threefold. First, the process of craniofacial development is largely conserved between mice and humans (Suzuki et al., 2016), which suggests that murine scRNA-seq data can be useful in terms of studying craniofacial development in the absence of human data. Second, the respective scRNA-seq samples were obtained at the time period of murine primary and secondary palate development (Miyake et al., 1996), thus increasing their suitability for studying CL/P candidate genes. Finally, research has shown that a large proportion of human embryonic scRNA-seq data from later developmental time points can be integrated with the MOCA data (Cao et al., 2020), providing further evidence for the transferability of developmental expression patterns. Although the MOCA scRNA-seq data are easily accessible via a comprehensive web browser, a systematic analysis in this setting is challenging. Of the 38 major cell clusters originally reported in MOCA, the present re-analysis identified a total of 31. This was probably attributable to differences in processing, since in the present study, the data were first split in accordance with embryonic day (in order to reduce the size of the data set to a computable level), followed by the performance of clustering. Nevertheless, as in the original MOCA publication, less diffuse clustering of some cell types was observed over the 5 day time-period, and a joint clustering of mesenchymal-like cell types was identified, such as chondrocyte progenitors, connective tissue progenitors, chondrocytes and osteoblasts, and jaw and tooth progenitors (commencing at E10.5). With regards to the facial dataset, the present analysis identified 25 clusters as opposed to 24 main clusters reported in the original publication. While we consider the numbers of clusters similar, we observed differences in cluster annotations. On one side, our re-analysis yielded several distinct cluster annotations for four clusters that were annotated as one cluster each in Li et al. This increased the number of clusters comprising those cells. On the other hand, we also failed to identify four of the 24 original clusters, including nasolacrimal groove and dental epithelium (see Supplementary Table S1). Investigating this further, we identified marker genes for these two clusters to be predominantly expressed in some of the cells of our ectodermal surface clusters and palatal epithelium, respectively (Supplementary Figure S5). Yet, these clusters did not split further into distinct clusters when using higher resolution clustering (*data not shown*). This divergence may be attributable to the fact that the present analysis involved a stricter filtering strategy, no cell cycle regression, and high-resolution clustering of all cells together without sub-clustering (as opposed to the original study that divided the data into ectoderm and mesenchyme first, and performed sub-

clustering at different resolutions individually). Together, this contributed to a lower absolute number of overall cells (7,249 compared to 7,893 from the original) and different clusters in our re-analysis.

Comparison of E11.5 transcriptome profiles between the MOCA and the facial data revealed substantial similarities at both the cell type and gene levels. For instance, red and white blood cells, endothelial cells, and Schwann cells represent distinct cell clusters that mapped at certain distances to the other clusters within the UMAP space. At the gene level, *Irf6*, *Tfap2a*, *Fras1*, *Cdh1,* and *Esrp1* exhibited similar expression patterns in epithelial cell types of both data sets. Additionally, *Tfap2a* showed expression in Schwann cell progenitors in both data sets. Together, these data suggest that the facial data set is a tissue-restricted, but high-resolution representation of the MOCA data at E11.5, and that collectively, the two datasets represent a valuable resource for genomics research into craniofacial development. However, caution is generally required when interpreting expression profiles from several scRNA-seq data sets, since scRNA-seq itself but also the combination of different sources have some limitations. These include differences in cell capture efficiency and transcript coverage, which may result in transcripts not being detected in all cells equally, and different enrichment strategies used in both studies. In addition, scRNA-seq data of tissues undergoing continuous processes during development, such as epithelial-to-mesenchymal transitions, only provide a snapshot of a possibly transient period of gene expression. Finally, varying sequencing depth adds to higher noise levels in scRNA-seq data compared to bulk RNA-seq data (Kolodziejczyk et al., 2015).

The expression patterns observed in the aggregated scRNA-seq data sets replicate previously reported and experimentally verified expression patterns. For instance, a previous study showed that *Irf6* is expressed in neural ectoderm and neural crest cells as early as E9.5 in murine embryonic development (Kousa et al., 2019). According to previous wet-lab data, *Irf6* is expressed in the ectoderm of the first and second pharyngeal arches, and in the palatal, lingual, maxillary, and mandibular epithelia, during the period E10.5–E13.5 (Kondo et al., 2002; Knight et al., 2006; Richardson et al., 2009; Goudy et al., 2013; Kousa and Schutte, 2016). In accordance, our analyses revealed the presence of *Irf6* expression in Schwann cell precursors, the palatal and olfactory epithelia, the ectodermal surface, the basal cells at the fusion zone, and the periderm in both, MOCA and facial data respectively. Of these, the highest expression was observed in the periderm and the basal cells at the fusion zone. Interestingly, only ~3% of the MOCA E11.5 epithelial cells expressed *Irf6*, as opposed to 40%–70% of those in the facial data set. This suggests that *Irf6*-expressing MOCA E11.5 epithelial cells might be derived from facial structures, while the epithelial cell cluster contains a substantial proportion of non-facial cells. Comparably, *Tfap2a* showed expression in the MOCA epithelial cells, as well as high expression levels in the facial ectodermal surface clusters and periderm. In addition, *Tfap2a* showed expression in Schwann cell precursor cells in both the MOCA and the facial data sets. Again, this expression pattern recapitulates existing data, since previous reports have demonstrated that in mice, *Tfap2a* is expressed in the ectoderm, cranial neural crest cells, the facial mesenchyme, nasal and oral epithelia, and the central and peripheral nervous system between E9–E13.5 (Mitchell et al., 1991; Chazaud et al., 1996; Moser et al., 1997). Previous studies have shown that *Esrp1* is expressed in the head region and epithelial cells, especially in cells of the ectodermal surface as early as E9.5 in mice (Warzecha et al., 2009; Revil and Jerome-Majewska, 2013; Bebee et al., 2015; Lee et al., 2020). Similarly, our data showed a broad expression of *Esrp1* in epithelial cells of both data sets with the highest expression in the periderm in the facial data set. Furthermore, the transcription factor *Foxe1* was found to be expressed in epithelial cells of embryonic mice starting at E9.5, both in our data sets and in previous studies (Zannini et al., 1997; Dathan et al., 2002; Moreno et al., 2009). In addition, studies have shown the keratin genes *Krt8* and *Krt18* to be expressed in single-layered epithelia in embryos (Jackson et al., 1980; Owens and Birgitte Lane, 2003; Moll et al., 2008). This is also confirmed by our data, as *Krt8* and *Krt18* showed expression in epithelial cells in both data sets. However, as expected, the strongest expression in the facial data set was found in the periderm and the palatal and olfactory epithelia. In contrast to the previously described genes, *Fgfr1* has been shown to be primarily expressed in mesenchymal cell types (Bachler and Neubüser, 2001), whis is also evident in our data, as *Fgfr1* was predominantly expressed in mesenchymal cell types. These similarities indicate that: (i) the data sets are reliable resources in the context of craniofacial development; and (ii) that expression patterns of genes that have not yet been experimentally validated may be characterized using scRNA-seq data.

In a first attempt to use these data in the context of CL/P, the present analyses identified two groups of CL/P candidate genes based on their expression in relevant facial cell types. Using predefined lists of CL/P candidate genes, the analyses identified distinct sets of genes that are predominantly expressed in either epithelial cells, or mesenchymal-like cells. Unsurprisingly, the first group included *Irf6*, *Tfap2a,* and *Esrp1*, which show similar expression in the six epithelial cell types of the facial data set, and which have been implicated in a regulatory network (Kousa et al., 2019; Carroll et al., 2020). A specific examination of the expression of the *Irf6-Grhl3-Tfap2a* genetic pathway revealed partial co-expression of *Irf6*, *Grhl3,* and *Tfap2a* within epithelial cells. This opens up the possibility that other CL/P candidate genes, which are among the marker genes of the *Irf6*+ epithelial cell types, or genes with an as yet unknown role in CL/P etiology, might also contribute to the *Irf6* regulatory network. We plan to follow up on this question in a future study, using more systematic co-expression network approaches (Dam et al., 2018). While the expression of *Irf6* in the periderm has already been established (Richardson et al., 2009; Richardson et al., 2014; Kousa et al., 2017), the scRNA-seq data suggest the presence of a specific sub-cell type in which *Irf6* and other CL/P candidate genes show co-expression, and that may contribute to the etiology of CL/P. Furthermore, the expression of CL/P candidate genes in adjacent facial cell types highlights CL/P candidate genes that might contribute to molecular communication between the different epithelial cell types, e.g., *Tpm1, Fras1, Krt7, Wnt7a, Rhpn2,* and *Sema3e* in the ectodermal surface clusters and periderm; and *Filip1l* in the ectodermal surface and cells adjacent to the ectodermal surface. These questions need to be addressed in the future using more sophisticated computational and experimental approaches, such as spatial transcriptomic analyses (Carangelo et al., 2022).

In a second application example, the MOCA data set was used to investigate potential differences in expressing cell types between syCL/P and nsCL/P candidate genes. In accordance with our hypothesis, syCL/P candidate genes were expressed in a larger number of cell types during the examined time period compared

to candidate genes for nsCL/P. Similar patterns were observed in the analysis of the gene lists for CHD. The AD syndromic CHD genes were expressed in significantly more cell types than the AR syndromic and the non-syndromic CHD genes (Supplementary Figure S2C). The average expression levels of the AD syndromic CHD genes were significantly higher than those of the non-syndromic CHD genes (Supplementary Figure S2D). While the precise reason for this effect requires further investigation, our analysis indicates the value of scRNAseq data in terms of the investigation of the different genetic architectures of CL/P subtypes.

In summary, the present study involved a re-analysis of previously published scRNA-seq data. We demonstrate the value of these data using several application examples. Our processed data sets are provided in Seurat object format as an easily accessible addition to the original data (see "Data availability statement"). This resource will facilitate functional approaches to the genomics of craniofacial development and disease.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, the processed data sets can be downloaded from https://beta.fastgenomics.org/p/Siewert_2023. Further inquiries can be directed to the corresponding author.

Author contributions

AS and KUL conceived and designed the study. BR and HD provided computational and infrastructural resources and assisted with the setup of the analytical environment. EM contributed genetic data. AS performed the data analysis and statistical testing, with the support of CK and JH. AS and KUL interpreted the data and drafted the first version of the manuscript. All authors have reviewed the final version of the article and approved its submission for publication.

Funding

KUL is supported by grants from the German Research Council (Deutsche Forschungsgemeinschaft (DFG), LU-1944/3-1).

Acknowledgments

The authors thank Christine Fischer for providing the mouse embryo graphics, Friederike David for technical support during data analysis, and Prof. Maximilian Billmann for helpful discussions.

Conflict of interest

BR and HD were employed by FASTGenomics (Comma Soft AG).

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2023.1091666/full#supplementary-material

References

Bachler, M., and Neubüser, A. (2001). Expression of members of the fgf family and their receptors during midfacial development. *Mech. Dev.* 100 (2), 313–316. doi:10.1016/S0925-4773(00)00518-9

Bartsocas, C. S., and Papas, C. V. (1972). Popliteal pterygium syndrome. Evidence for a severe autosomal recessive form. *J. Med. Genet.* 9 (2), 222–226. doi:10.1136/jmg.9.2.222

Bebee, T. W., Juw Won Park, K. I., Alex, M., Xing, Y., and Carstens, R. P. (2015). The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. *ELife* 4, 1–27. doi:10.7554/eLife.08954

Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Rubini, M., Baluardo, C., et al. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat. Genet.* 41 (4), 473–477. doi:10.1038/ng.333

Bishop, M. R., Diaz Perez, K. K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., et al. (2020). Genome-wide enrichment of de novo coding mutations in orofacial cleft trios. *Am. J. Hum. Genet.* 107 (1), 124–136. doi:10.1016/j.ajhg.2020.05.018

Butali, A., Mossey, P. A., Adeyemo, W. L., Eshete, Me. A., Gowans, L. J. J., Busch, T. D., et al. (2019). Genomic analyses in african populations identify novel risk loci for cleft palate. *Hum. Mol. Genet.* 28 (6), 1038–1051. doi:10.1093/hmg/ddy402

Cao, J., Diana, R., Pliner, H. A., Paul, D. K., Deng, M., Riza, M., et al. (2020). A human cell Atlas of fetal gene expression. *Sci. (New York, N.Y.)* 370 (6518), 7721. doi:10.1126/science.aba7721

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., et al. (2019). The single-cell transcriptional landscape of mammalian Organogenesis. *Nature* 566 (7745), 496–502. doi:10.1038/s41586-019-0969-x

Carangelo, G., Magi, A., and Semeraro, R. (2022). From multitude to singularity: An up-to-Date overview of ScRNA-seq data generation and analysis. *Front. Genet.* 13, 994069–994116. doi:10.3389/fgene.2022.994069

Carroll, S. H., Macias Trevino, C., Li, E. B., Kawasaki, K., Myers, N., Hallett, S. A., et al. (2020). An irf6-esrp1/2 regulatory Axis controls midface Morphogenesis in vertebrates. *Dev. Camb.* 147, dev194498. doi:10.1242/dev.194498

Chazaud, C., Oulad-Abdelghani, M., Bouillet, P., Décimo, D., Chambon, P., and Dollé, P. (1996). AP-2.2, a novel gene related to AP-2, is expressed in the forebrain, limbs and face during mouse embryogenesis. *Mech. Dev.* 54 (1), 83–94. doi:10.1016/0925-4773(95)00463-7

Cox, L. L., CoxMoreno, T. C. L. M. U., Zhu, Y., Richter, C. T., Nidey, N., Standley, J. M., et al. (2018). Mutations in the epithelial cadherin-P120-catenin complex cause mendelian non-syndromic cleft lip with or without cleft palate. *Am. J. Hum. Genet.* 102 (6), 1143–1157. doi:10.1016/j.ajhg.2018.04.009

Dam, S. V., Võsa, U., Franke, L., and Pedro de Magalhães, J. (2018). Gene Co-expression analysis for functional classification and gene-disease predictions. *Briefings Bioinforma.* 19 (4), 575–592. doi:10.1093/bib/bbw139

Dathan, N., Parlato, R., Rosica, A., De Felice, M., and Di Lauro, R. (2002). Distribution of the titf2/foxe1 gene product is consistent with an important role in the development

of foregut endoderm, palate, and hair. *Dev. Dyn.* 224 (4), 450–456. doi:10.1002/dvdy.10118

Evans, D. G. R., Ladusans, E. J., Rimmer, S., Burnell, L. D., Thakker, N., and Farndon, P. A. (1993). Complications of the naevoid basal cell carcinoma syndrome: Results of a population based study. *J. Med. Genet.* 30 (6), 460–464. doi:10.1136/jmg.30.6.460

Fraser, G. R. (1962). Our genetical 'load': A review of some aspects of genetical variation. *Ann. Hum. Genet.* 25, 387–415. doi:10.1111/j.1469-1809.1962.tb01774.x

Ghoumid, J., Stichelbout, M., Frenois, F., Lejeune-Dumoulin, S., Alex-Cordier, M., Lebrun, M., et al. (2017). Blepharocheilodontic syndrome is a CDH1 pathway-related disorder due to mutations in CDH1 and CTNND1. *Genet. Med.* 19 (9), 1013–1021. doi:10.1038/gim.2017.11

Goudy, S., Angel, P., Jacobs, B., Hill, C., Mainini, V., Smith, A. L., et al. (2013). Cell-autonomous and non-cell-autonomous roles for Irf6 during development of the tongue. *PLoS ONE* 8 (2), e56270. doi:10.1371/journal.pone.0056270

Hao, Y., Stephanie, H., Erica Andersen-Nissen, W. M. M., Zheng, S., Butler, A., Lee, M. J., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell.* 184 (13), 3573–3587. doi:10.1016/j.cell.2021.04.048

Jackson, B. W., Christine, G., Erika, S., Kurt, B., Werner, W., and Karl, I. (1980). Formation of cytoskeletal elements during mouse embryogenesis: Intermediate filaments of the cytokeratin type and desmosomes in preimplantation embryos. *Differentiation* 17 (1–3), 161–179. doi:10.1111/j.1432-0436.1980.tb01093.x

Ji, Y., Garland, M. A., Sun, B., Zhang, S., Reynolds, K., McMahon, M., et al. (2020). Cellular and developmental basis of orofacial clefts. *Birth Defects Res.* 112 (19), 1558–1587. doi:10.1002/bdr2.1768

Kimonis, V. E., Goldstein, A. M., Pastakia, B., Yang, M. L., Kase, R., Digiovanna, J. J., et al. (1997). Clinical manifestations in 105 persons with nevoid basal cell carcinoma syndrome. *Am. J. Med. Genet.* 69 (3), 299–308. doi:10.1002/(SICI)1096-8628(19970331)69:3<299::AID-AJMG16>3.0.CO;2-M

Kimonis, Vi. E., Singh, K. E., Zhong, R., Pastakia, B., Digiovanna, J. J., and Bale, S. J. (2013). Clinical and radiological features in young individuals with nevoid basal cell carcinoma syndrome. *Genet. Med.* 15 (1), 79–83. doi:10.1038/gim.2012.96

Knight, A. S., Schutte, B. C., Jiang, R., and Dixon, M. J. (2006). Developmental expression analysis of the mouse and chick orthologues of IRF6: The gene mutated in van der Woude syndrome. *Dev. Dyn.* 235 (5), 1441–1447. doi:10.1002/dvdy.20598

Kolodziejczyk, A. A., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and Biology of single-cell RNA sequencing. *Mol. Cell.* 58 (4), 610–620. doi:10.1016/j.molcel.2015.04.005

Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, Al. S., Watanabe, Y., et al. (2002). Mutations in IRF6 cause van der Woude and popliteal pterygium syndromes. *Nat. Genet.* 32 (2), 285–289. doi:10.1038/ng985

Kousa, Y. A., Roushangar, R., Patel, N., Walter, A., Marangoni, P., Krumlauf, R., et al. (2017). IRF6 and SPRY4 signaling interact in periderm development. *J. Dent. Res.* 96 (11), 1306–1313. doi:10.1177/0022034517719870

Kousa, Y. A., and Schutte, B. C. (2016). Toward an orofacial gene regulatory network. *Dev. Dyn.* 245 (3), 220–232. doi:10.1002/dvdy.24341

Kousa, Y. A., Zhu, H., Fakhouri, W. D., Lei, Y. A. K., Roushangar, R. R., Patel, N. K., et al. (2019). The tfap2a-IRF6-GRHL3 genetic pathway is conserved in neurulation. *Hum. Mol. Genet.* 28 (10), 1726–1737. doi:10.1093/hmg/ddz010

Lee, S. K., Sears, M. J., Zhang, Z., Hong, L., Salhab, I., Krebs, P., et al. (2020). Cleft lip and cleft palate in Esrp1 knockout mice is associated with alterations in epithelial-mesenchymal crosstalk. *Dev. Camb.* 147 (21), dev187369. doi:10.1242/dev.187369

Leslie, E. J., Carlson, J. C., Eleanor Feingold, J. R. S., George, W., Laurie, C. A., Jain, D., et al. (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p 24.2, 17q23 and 19q13. *Hum. Mol. Genet.* 25 (13), 2862–2872. doi:10.1093/hmg/ddw104

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Butali, A., Carmen, J., Castilla, E. E., et al. (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Hum. Genet.* 136 (3), 275–286. doi:10.1007/s00439-016-1754-7

Li, H., Jones, K. L., Hooper, J. E., and Williams, T. (2019). The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. *Dev. Camb.* 146 (12), dev174888. doi:10.1242/dev.174888

Li, M. J., JiaS, Y., Zhu, S., Shi, B., and Zhong, L. (2022). Targeted Re-sequencing of the 2p21 locus identifies non-syndromic cleft lip only novel susceptibility gene ZFP36L2. *Front. Genet.* 13, 802229. doi:10.3389/fgene.2022.802229

Lin-Shiao, E., Lan, Y., Welzenbach, J., Alexander, K. A., Zhang, Z., Knapp, M., et al. (2019). P63 establishes epithelial enhancers at critical craniofacial development genes. *Sci. Adv.* 5 (5), eaaw0946–15. doi:10.1126/sciadv.aaw0946

Ludwig, K. U., Mangold, E., Herms, S., Nowak, S., Paul, A., Becker, J., et al. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Palate Identify Six. New Risk Loci* 44 (9), 968–971. doi:10.1038/ng.2360

Mangold, E., Anne, C., Hoebel, A. K., and Gültepe, P. (2016). Sequencing the GRHL3 coding region reveals rare truncating mutations and a common susceptibility variant for nonsyndromic cleft palate. *Am. J. Hum. Genet.* 98 (4), 755–762. doi:10.1016/j.ajhg.2016.02.013

Mangold, E., Ludwig, K. U., and Nöthen, M. M. (2011). Breakthroughs in the genetics of orofacial clefting. *Trends Mol. Med.* 17 (12), 725–733. doi:10.1016/j.molmed.2011.07.007

Milunsky, J. M., Maher, T. A., Roberts, A. E., Stalker, H. J., Zori, R. T., Burch, M. N., et al. (2008). TFAP2A mutations result in branchio-oculo-facial syndrome. *Am. J. Hum. Genet.* 82 (5), 1171–1177. doi:10.1016/j.ajhg.2008.03.005

Mitchell, P. J., Timmons, P. M. J. M., Tjian, R., and Rigby, P. W. (1991). Transcription factor AP-2 is expressed in neural crest cell lineages during mouse embryogenesis. *Genes. Dev.* 5 (1), 105–119. doi:10.1101/gad.5.1.105

Miyake, T., Cameron, A. M., and Hall, B. K. (1996). Detailed staging of inbred C57BL/6 mice between Theiler's [1972] stages 18 and 21 (11–13 days of gestation) based on craniofacial development. *J. Craniofac. Genet. Dev. Biol.* 16, 1–31.

Moll, R., Divo, M., and Lutz, L. (2008). The human keratins: Biology and pathology. *Histochem. Cell. Biol.* 129 (6), 705–733. doi:10.1007/s00418-008-0435-6

Moreno, L. M., Maria Adela Mansilla, S. A. B., Cooper, M. E., Busch, T. D., Johnson, M. K., Busch, T. D., et al. (2009). FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Hum. Mol. Genet.* 18 (24), 4879–4896. doi:10.1093/hmg/ddp444

Moser, M., Rüschoff, J., and Buettner, R. (1997). Comparative analysis of AP-2α and AP-2β gene expression during murine embryogenesis. *Dev. Dyn.* 208 (1), 115–124. doi:10.1002/(SICI)1097-0177(199701)208:1<115::AID-AJA11>3.0.CO;2-5

Murray, J. C. (2002). Gene/environment causes of cleft lip and/or palate. *Clin. Genet.* 61 (9), 248–256. doi:10.1034/j.1399-0004.2002.610402.x

Nees, S. N., and Chung, W. K. (2020). Genetic basis of human congenital heart disease. *Cold Spring Harb. Perspect. Biol.* 12 (9), 0367499–a36840. doi:10.1101/cshperspect.a036749

Niemann, S., Zhao, C., Pascu, F., Stahl, U., Aulepp, U., Lee, N., et al. (2004). Homozygous WNT3 mutation causes tetra-amelia in a large consanguineous family. *Am. J. Hum. Genet.* 74 (3), 558–563. doi:10.1086/382196

Owens, D. W., and Birgitte Lane, E. (2003). The quest for the function of simple epithelial keratins. *BioEssays* 25 (8), 748–758. doi:10.1002/bies.10316

Revil, T., and Jerome-Majewska, L. A. (2013). During embryogenesis, Esrp1 expression is restricted to a subset of epithelial cells and is associated with splicing of a number of developmentally important genes. *Dev. Dyn.* 242 (3), 281–290. doi:10.1002/dvdy.23918

Richardson, R. J., Nigel, L., Hammond, P. A. C., Saloranta, C., Nousiainen, H. O., Salonen, R., et al. (2014). Periderm prevents pathological epithelial adhesions during embryogenesis. *J. Clin. Investigation* 124 (9), 3891–3900. doi:10.1172/JCI71946

Richardson, R. J., Dixon, J., Jiang, R., and Dixon, M. J. (2009). Integration of IRF6 and Jagged2 signalling is essential for controlling palatal adhesion and fusion competence. *Hum. Mol. Genet.* 18 (14), 2632–2642. doi:10.1093/hmg/ddp201

Scholz, C., Biernat, P., Becker, M., Bassler, K., Günther, P., Balfer, J., et al. (2018). FASTGenomics: An analytical ecosystem for single-cell RNA sequencing data partial. BioRxiv.

Shao, X., Liao, J., Lu, X., Xue, R., Ni, A., and Fan, X. (2020). ScCATCH: Automatic annotation on cell types of clusters from single-cell RNA sequencing data. *IScience* 23 (3), 100882. doi:10.1016/j.isci.2020.100882

Suzuki, A., Sangani, D. R., Ansari, A., and Iwata, J. (2016). Molecular mechanisms of midfacial developmental defects. *Dev. Dyn.* 245 (3), 276–293. doi:10.1002/dvdy.24368

Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B., and Carstens, R. P. (2009). ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell.* 33 (5), 591–601. doi:10.1016/j.molcel.2009.01.025

Welzenbach, J., Hammond, N. L., Nikolić, M., Thieme, F., Ishorst, N., Leslie, E. J., et al. (2021). Integrative approaches generate insights into the architecture of non-syndromic cleft lip ± cleft palate. *Hum. Genet. Genomics Adv.* 2 (3), 100038–100114. doi:10.1016/j.xhgg.2021.100038

Woude, A. V. D. (1954). Fistula labii inferioris congenita and its association with cleft lip and palate. *Am. J. Hum. Genet.* 6 (2), 244–256.

Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., et al. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat. Commun.* 8 (2), 14364–14411. doi:10.1038/ncomms14364

Zannini, M., Avantaggiato, V., Biffali, E., Sato, K., Pischetola, M., Taylor, B. A., et al. (1997). TTF-2, a new forkhead protein, shows a temporal expression in the developing thyroid which is consistent with a role in controlling the onset of differentiation. *EMBO J.* 16 (11), 3185–3197. doi:10.1093/emboj/16.11.3185

4.2 Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of *de novo* mutations

Hanna K. Zieger, Leonie Weinhold, Axel Schmidt, Manuel Holtgrewe, Stefan A. Juranek, **Anna Siewert**, Annika B. Scheer, Frederic Thieme, Elisabeth Mangold, Nina Ishorst, Fabian U. Brand, Julia Welzenbach, Dieter Beule, Katrin Paeschke, Peter M. Krawitz, and Kerstin U. Ludwig

Human Genetics and Genomics Advances 1:100166 (2023)

https://doi.org/10.1016/j.xhgg.2022.100166

HGG Advances

ARTICLE

Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of *de novo* mutations

Hanna K. Zieger,[1] Leonie Weinhold,[2] Axel Schmidt,[1] Manuel Holtgrewe,[3] Stefan A. Juranek,[4] Anna Siewert,[1] Annika B. Scheer,[1] Frederic Thieme,[1] Elisabeth Mangold,[1] Nina Ishorst,[1] Fabian U. Brand,[5] Julia Welzenbach,[1] Dieter Beule,[3,6] Katrin Paeschke,[4] Peter M. Krawitz,[2] and Kerstin U. Ludwig[1,*]

Summary

Non-syndromic cleft lip with/without cleft palate (nsCL/P) is a highly heritable facial disorder. To date, systematic investigations of the contribution of rare variants in non-coding regions to nsCL/P etiology are sparse. Here, we re-analyzed available whole-genome sequence (WGS) data from 211 European case-parent trios with nsCL/P and identified 13,522 *de novo* mutations (DNMs) in nsCL/P cases, 13,055 of which mapped to non-coding regions. We integrated these data with DNMs from a reference cohort, with results of previous genome-wide association studies (GWASs), and functional and epigenetic datasets of relevance to embryonic facial development. A significant enrichment of nsCL/P DNMs was observed at two GWAS risk loci (4q28.1 ($p = 8 \times 10^{-4}$) and 2p21 ($p = 0.02$)), suggesting a convergence of both common and rare variants at these loci. We also mapped the DNMs to 810 position weight matrices indicative of transcription factor (TF) binding, and quantified the effect of the allelic changes *in silico*. This revealed a nominally significant overrepresentation of DNMs ($p = 0.037$), and a stronger effect on binding strength, for DNMs located in the sequence of the core binding region of the TF Musculin (MSC). Notably, MSC is involved in facial muscle development, together with a set of nsCL/P genes located at GWAS loci. Supported by additional results from single-cell transcriptomic data and molecular binding assays, this suggests that variation in MSC binding sites contributes to nsCL/P etiology. Our study describes a set of approaches that can be applied to increase the added value of WGS data.

Introduction

Non-syndromic cleft lip with/without cleft palate (nsCL/P) is the most frequent form of orofacial clefting (OFC), with an estimated prevalence of 1 in 1,000 European newborns.[1] Depending on severity, nsCL/P treatment requires multidisciplinary approaches, including repeated surgeries, throughout childhood and adolescence. Together with an increased life-time risk for morbidity and mortality,[2] nsCL/P represents a major burden for affected individuals and their families.

NsCL/P has a multifactorial etiology, and estimates from twin studies suggest a heritability of ~90%.[3] Recent genome-wide association studies (GWASs) have identified common risk variants at 45 genomic loci, which explain about 30% of phenotypic variance in Europeans.[4] Research suggests that further types of genetic variation may also contribute to disease risk, including variants from the low-frequency part of the allelic spectrum. For example, previous studies have identified private and rare risk variants for nsCL/P in genes underlying orofacial cleft syndromes within multiplex families,[5] in genes involved in epithelial cell adhesion processes,[6] and in genes located within GWAS loci.[7–10] In a recent multiethnic study of several hundred case-parent trios of OFC (Bishop et al.),[11] potentially causal *de novo* mutations (DNMs) in protein-coding regions were investigated using data from whole-genome sequencing (WGS). The cohort included individuals with cleft lip with/without cleft palate (CL/P), including its subtypes cleft lip only (CLO) as well as cleft lip and palate (CLP), and cleft palate only (CPO). In that study, the authors identified a cohort-wide enrichment of loss of function (LoF) DNMs, in particular in genes expressed in human neural crest cells (hNCCs). At the individual gene level, this study also implicated *TFAP2A* (MIM: 107580), *IRF6* (MIM: 607199), and *ZFHX4* (MIM: 606940) in OFC etiology.[11]

To date, most analyses of systematic sequencing data (including Bishop et al.) have been limited to protein-coding regions, mainly because of the comparable ease of functional annotation and etiological interpretation for coding variants. In contrast, few data are available concerning the contribution of rare variants or DNMs located in non-coding regions. Evidence that non-coding variants are involved in nsCL/P has been generated by studies that identified causal non-coding mutations in individual pedigrees,[10,12,13] and reports of a burden of low-frequency variants in non-coding enhancer regions that are active in

[1]Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn 53127, Germany; [2]Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn 53127, Germany; [3]Core Unit Bioinformatics, Berlin Institute of Health, Berlin 10117, Germany; [4]Department of Oncology, Hematology and Rheumatology, University Hospital Bonn, Bonn 53127, Germany; [5]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn 53127, Germany; [6]Max Delbrück Center for Molecular Medicine, Berlin 13125, Germany
*Correspondence: kerstin.ludwig@uni-bonn.de
https://doi.org/10.1016/j.xhgg.2022.100166.

developing craniofacial tissue.[14,15] The aim of the present study was to identify etiologically relevant DNMs for nsCL/P, with a focus on strategies to prioritize DNMs in non-coding regions.

Material and methods

This study used prior published data, no human or animal subjects were involved. Respective datasets were analyzed upon approved data access and following the criteria laid out in the respective data use agreements in the NIH database of Genotypes and Phenotypes (dbGaP). Informed consent and ethical approval were obtained by the investigators of the original studies. The molecular and computational studies did not involve any human material. All procedures followed biological safety and ethics standards.

Subjects and data resources

WGS raw sequence and phenotypic data for 1,236 individuals from a European OFC cohort were retrieved from the Gabriella Miller Kids First (GMKF) Project, upon approved access (section "Web resources"). Based on available pedigree information, 220 complete parent-offspring pairs ("trios") containing both unaffected parents and a child with nsCL/P were identified. Additionally, a set of 330 trios with children being affected by Ewing sarcoma (ES) was obtained from GMFK. This cohort was used as a non-cleft reference (NCR) cohort. Further information can be found in the supplemental methods.

WGS data analysis and variant calling

For each individual, WGS reads were aligned to GRCh37, and variant calling was performed using both Unified Genotyper and Haplotype Caller. To generate a high-quality variant DNM call set, data processing required the complete absence of reads in any parent, and support of variant calls by both calling algorithms (supplemental methods). All DNMs were annotated with information (1) on frequency (gnomAD v3.1, all populations), (2) on genomic location (exonic, intronic, intergenic; based on GENCODE Basic gene annotation version33.hg19), and (3) with each of six *in silico* prediction scores that are applicable to both non-coding and coding variants: CADD,[16] ReMM,[17] FATHMM,[18] DANN,[19] LINSIGHT,[20] and ncER[21] (supplemental methods). No general frequency filter was applied (Figure S1). As our nsCL/P cohort represents a subcohort of Bishop et al. that was analyzed using a different quality control (QC) and variant calling pipeline, coding DNMs were compared between both studies, based on available information (Table S3 by Bishop et al., participant IDs provided by GMKF) and annotations provided by the Ensembl Variant Effect Predictor[22] (VEP; section "web resources").

The statistical comparison of DNM distribution between nsCL/P and NCR included the average number of DNMs per sample (Mann-Whitney U (MWU) test for total DNMs and subgroups of exonic, intronic, and intergenic DNMs), the distribution of *in silico* prediction scores for nsCL/P and NCR DNMs, and the proportion of DNMs with *in silico* prediction scores over individual or combined thresholds (supplemental methods).

Analysis of DNM enrichment in genomic features

To study the enrichment of DNMs across the entire genome, diverse genomic datasets were retrieved. For each of those datasets, DNM enrichment was calculated using the R package FunciVar,[23] which compares inter-cohort enrichment probabilities for functional elements using a Bayesian approach (see FunciVar in section "web resources," supplemental methods). The datasets included genome-wide maps of eight chromatin states from hNCCs,[24] cranial neural crest cells (cNCCs),[25] and human facial embryonic tissues,[26] which had been aggregated in a previous study by our group.[4] Furthermore, general genomic features with *a priori* evidence for functional relevance or evolution were included; i.e., (1) 4,307 evolutionarily highly conserved non-coding elements (CNEs) based on a prior publication,[27] and (2) 1,570 enhancer regions from the VISTA enhancer browser[28] (supplemental methods).

Analysis of topologically associating domains

To detect local enrichments of non-coding DNMs independent of genomic features (comparable with gene-burden tests for protein-coding variants), DNMs were combined based on their location within regulatory units; i.e., topologically associating domains (TADs). Positional data were retrieved for 2,991 TADs from human embryonic stem cells, as described elsewhere,[4] and enrichment of DNMs in TADs was tested using FunciVar (supplemental methods). Given the considerable burden of multiple testing with regard to the present sample size, we additionally defined a set of 45 candidate TADs on the basis of recent GWAS results, as previously described[4] ($TADs_{GWAS}$, Table S1).

Analysis of DNMs in TF binding sites

Position weight matrix (PWM) information representing 810 transcription factor binding site (TFBS) motifs was retrieved from JASPAR2020.[29] Using a modified version of a previously published pipeline (see denovoLOBGOB, sections "Web resources," "data and code availability"), changes in transcription factor (TF) binding between reference and alternative alleles were qualitatively predicted and quantified for each DNM (after excluding insertions/deletions (indels); n = 28,773 DNMs). Statistical analyses of individual PWMs were performed to determine (1) differences in how frequently a specific PWM matches the genomic region around the DNMs (Fisher's exact test), and (2) quantitative differences in predicted binding strength (MWU test). For the latter, for each DNM, the effect of the variant allele was calculated as described above, and the difference from the reference allele was determined as an absolute change of binding. Then, absolute change values were combined for all DNMs of one PWM and compared between the two cohorts. In addition, for each analysis (1) and (2), log2-fold changes (log2FC) between nsCL/P and NCR were calculated. Further information can be found in the supplemental methods.

Single-cell expression data

Single-cell expression data obtained from murine embryos were downloaded from (1) the Mouse Organogenesis Cell Atlas (MOCA), which includes a time series of developmental organogenesis from E9.5 to E13.5 (section "Web resources"); and (2) the lambdoidal junction at day E11.5, which represents the time point for the fusing of facial structures.[30] Both datasets were re-analyzed using a joint in-house computational pipeline (supplemental methods).

Electrophoretic mobility shift assays

For each of the DNMs observed within MSC binding sites, gain or loss of binding was predicted based on the allelic change within the motif: gain of binding (if PWM-ref < PWM-alt), loss of binding (PWM-ref > PWM-alt), and silent effects (PWM-ref = PWM-alt).

Table 1. Distribution of DNMs in nsCL/P and NCR trios

	nsCL/P	NCR	Combined
Total DNMs	13,522	17,968	31,490
SNVs	12,335	16,438	28,773
Small insertions/deletions	1,187	1,530	2,717
Protein-coding DNMs[a]	222 (1.05)[c]	338 (1.19)[c]	560
LoF DNMs[b]	22 (0.10)[c]	19 (0.07)[c]	41
Nonsense DNMs	10	11	21
Frameshift DNMs	12	8	20
Missense DNMs	129 (0.61)[c]	246 (0.87)[c]	375
Synonymous DNMs	71 (0.34)[c]	73 (0.26)[c]	144

DNMs, *de novo* mutations; nsCL/P, non-syndromic cleft lip with/without cleft palate; NCR, non-cleft reference cohort; LoF, loss of function.
[a]Exonic DNMs based on GENCODE Basic gene annotation version33.hg19, including non-coding parts of gene sequences (e.g., 3′/5′ UTRs).
[b]Effect combinations from Variant Effect Predictor output were reduced to classes (see Table S4 for grouped effect names). LoF DNMs include nonsense and frameshift DNMs.
[c]In brackets: relative frequency of this type of DNM in the respective cohort.

Then, five candidate binding sites were selected from the set of DNMs; i.e., two motifs located at nsCL/P DNMs with either the strongest loss (chromosome [chr.] 6, chr. 10) or strongest gain (chr. 7, chr. 16), and the motif with the strongest predicted binding change by DNM in NCR (chr. 5; Table S2). For each of the five candidate binding sites of MSC, the genomic context around the DNM (i.e., an additional 20 bp up- and downstream) was retrieved. Each target oligonucleotide was designed with the respective duplex reference and alternative motif, and each contained p^{32} marks at the 5′ end of the top strand. Following cloning of MSC into the pET-28a vector, expression in *Escherichia coli*, and purification, the protein was incubated with binding buffer and oligonucleotides, for 30 min. Then 10 nM DNA was incubated with five different concentrations of MSC (range 0–1 μM). Binding effects were monitored according to the presence of protein-oligo dimers at predicted molecular size on native gels, and potential allele-specific effects were indicated by gel mobility changes (supplemental methods, all tested sequences in Table S2). All analyses were performed in triplicate.

Results

High-confidence variant set of coding and non-coding DNMs

After sample- and variant QC (Figures S2, S3, and S4), the final dataset contained 211 nsCL/P trios (52 of which were CLO, and 159 CLP; Figures S5 and S6), 284 NCR trios, and 31,490 autosomal DNMs (13,522 in nsCL/P; 17,968 in NCR; Table 1). Among those, 28,773 DNMs were single-nucleotide variants (SNVs), and 2,717 were small indels. Sixteen DNMs were recurrent (four within nsCL/P, seven within NCR, and five were observed in both cohorts; Table S3). Overall, an average of 63.6 autosomal DNMs was observed per trio, consistent with expectations.[31] No significant difference in the average number of DNMs was observed between nsCL/P and NCR trios (64.1 versus 63.3; $p = 0.47$; Figure S7), and both cohorts showed a similar distribution of DNMs across exonic, intronic, and intergenic regions (Figure 1A).

Within the nsCL/P cohort, 222 of the exonic DNMs mapped within protein-coding sequences according to VEP (Tables 1, S4, and S5; supplemental methods). This included 22 LoF (12 frameshift, 10 nonsense), 129 missense (together denoted as protein-altering DNMs), and 71 synonymous variants. No splice site DNM was observed. Notably, 159 of the 222 coding DNMs were previously reported by Bishop et al. (=71.6%, supplemental methods). This indicates convergence of the identified DNMs between both studies, taking into account the differences in variant calling pipelines and quality parameters. An aggregation of all coding DNMs of this study and the study by Bishop et al. can be found in Table S6.

Identification of deleterious variants in craniofacial genes

We next annotated each of the 31,490 DNMs with six *in silico* prediction scores (i.e., CADD, ReMM, FATHMM, DANN, LINSIGHT, and ncER). Comparison of score distributions did not reveal conclusive differences between nsCL/P and NCR (Figures 1B, S8, S9, and S10; Tables S7, S8, S9, S10, S11, S12, S13, and S14), and filtering for DNMs with CADD $\geq$ 20 did not show a significant difference between cohorts ($p = 0.18$, 144 DNMs in nsCL/P [1.06%], 226 DNMs in the NCR cohort [1.26%]; Table S15). Notably, DNMs in numerous craniofacial genes, such as *WNT4* (MIM: 603490),[32,33] *ALPI* (MIM: 171740),[34] and *MYO10* (MIM: 601481)[35–37] were observed with high CADD scores of $\geq$30 in nsCL/P. In addition, one DNM (CADD score of 45) was observed in *PLEKHA6* (MIM: 607771), which is a paralog of *PLEKHA7* (MIM: 612686). Pathogenic variants in *PLEKHA7* were reported in a previous investigation of multiply affected nsCL/P families[6]; thereby, this result further supports the role of the PLEKHA-family in nsCL/P etiology.

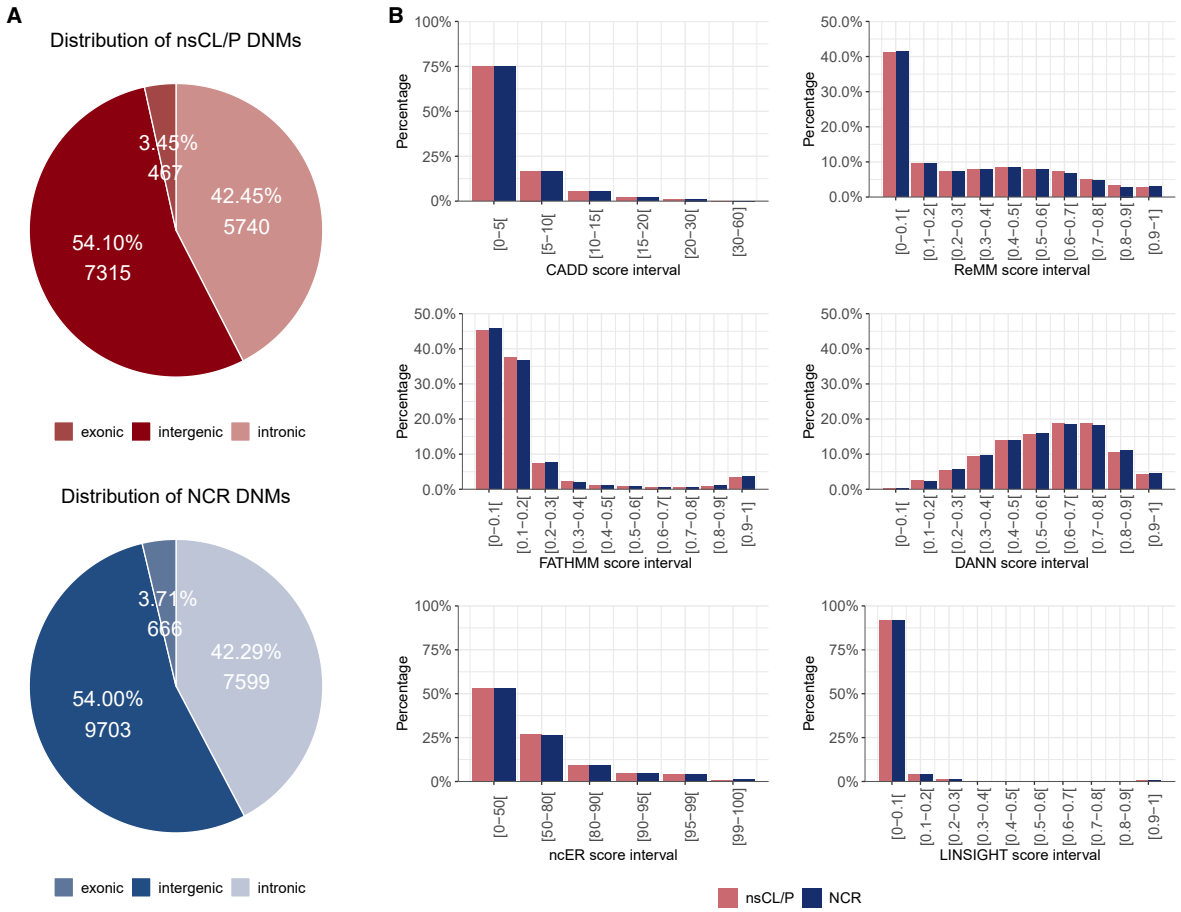


Figure 1. Comparative analyses of *de novo* mutations
(A) *De novo* mutations (DNMs) observed in non-syndromic cleft lip with/without cleft palate (nsCL/P) case-parent trios (red) and NCR trios (blue) were annotated according to genomic location (i.e., exonic/intronic/intergenic). Exonic DNMs were defined based on exons of protein-coding genes in the GENCODE Basic gene annotation version33.hg19, including non-coding parts of gene sequences (e.g., 3′/5′ UTRs). DNMs were equally distributed between the two cohorts.
(B) DNMs were annotated with each of six distinct *in silico* prediction scores, and their distribution was compared between the two cohorts. No significant differences were found.

Limited evidence for enrichment of non-coding DNMs in genomic features

We first tested the hypothesis that DNMs are significantly enriched in epigenetic and functional datasets of relevance to embryonic facial development. No analysis-wide enrichment was observed, with the exception of a nominal significant finding in bivalent/poised transcription start sites and bivalent enhancers of Carnegie stage 15 of human facial embryonic tissue[26] (74 DNMs [0.55%; Table S16] in nsCL/P versus 68 DNMs in the NCR cohort [0.38%], $p = 0.03$; Figure 2A; Table S17). While this enrichment is noteworthy, the failure of reaching robust levels of statistical evidence precludes a conclusive statement.

No enrichment was observed for 34 nsCL/P DNMs that mapped to any of 4,307 CNEs (Figure 2B, 15 in nsCL/P versus 19 in NCR cohort; Tables S18, S19, and S20; $p = 0.88$). Regarding the 40 DNMs mapping to VISTA enhancers, again, no significant difference was observed between the nsCL/P and NCR cohorts (14 versus 26; $p = 0.31$; Tables S21 and S22). This finding remained unchanged when DNMs were grouped for tissue-specific effects (activity in 16 of 23 different tissue types; Figure 2B; Table S23). Furthermore, no nsCL/P DNM was localized in both a CNE and a VISTA enhancer.

Convergence of non-coding DNMs at two GWAS risk loci

As TADs are considered the general regulatory units of the genome,[38] the aggregation of DNMs within its boundaries provides a systematic approach to aggregate DNMs with similar mechanistic effects. Based on the overall variant dataset, 29,629 DNMs were unambiguously mapped within 2,961 individual TADs (supplemental methods). While there was no test-wide significant difference between nsCL/P and NCR in terms of enrichment or depletion of DNMs in any of these TADs, we observed that 174 of the individual TADs showed a nominally significant

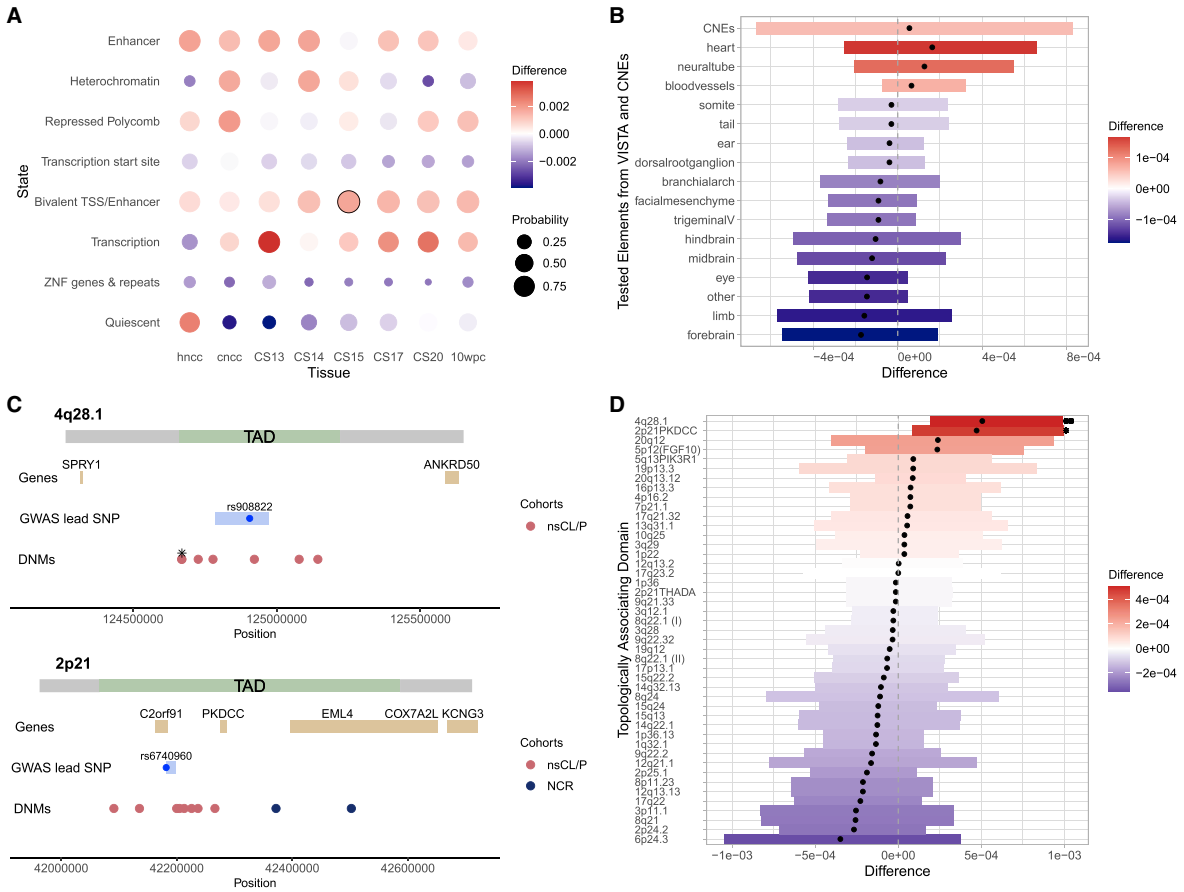


Figure 2. Enrichment of non-syndromic cleft lip with/without cleft palate *de novo* mutations in genomic candidate regions
(A) DNMs were mapped in eight chromatin states derived from human neural crest cells (hNCCs), cranial neural crest cells (cNCCs), and human embryonic facial tissue. FunciVar enrichment results are indicated by dot color. Dot sizes illustrate enrichment probabilities (increasing values represent increased statistical significance), and significant findings are encircled.
(B) Non-coding elements with previous evidence for functional relevance were retrieved from conserved non-coding elements (CNEs) and enhancer activity assays from VISTA (n=16 tissues). DNMs mapping to these regions were tested for n enrichment in nsCL/P using FunciVar, similar to (A), and enrichment was depicted with their respective 95% credible interval (dots indicate median). The gray dashed line indicates a difference of zero.
(C) DNMs were mapped within boundaries of topologically associating domains (TADs), and a subset of 45 TADs was defined based on the presence of associated common nsCL/P risk variants ($TADs_{GWAS}$). Two loci (4q28.1, $2p21_{PKDCC}$, see panel D) carried significantly more DNMs in nsCL/P. TAD boundaries are highlighted in green, with surrounding regions in gray. Gene locations are shown in yellow, together with GWAS-SNPs (dot) and GWAS credible SNP regions (bar) in blue. The positions of DNMs are indicated in red for nsCL/P and dark blue for NCR cohort. Two superimposed DNMs at 4q28.1 are indicated by an asterisk (*).
(D) Same graphical depiction as in (B), except for the TADs located at the 45 nsCL/P GWAS risk loci. Nominal significant p values are indicated with an asterisk (*), and p values significant after correction for 45 tests are indicated by a double asterisk (**).

enrichment (n = 98) or depletion (n = 76) of DNMs in nsCL/P compared with NCR (Table S24). Restricting the analysis to 45 $TADs_{GWAS}$, we observed 544 DNMs in total (221 nsCL/P versus 323 NCR), with two $TADs_{GWAS}$ showing significant enrichment of DNMs in nsCL/P; i.e., $2p21_{PKDCC}$[39] and 4q28.1[40] (Figure 2C; Tables S25 and S26). At the 4q28.1 locus, seven DNMs were observed in seven different individuals with nsCL/P, while no DNM in this region was observed in the NCR cohort ($p = 8 \times 10^{-4}$). At the $2p21_{PKDCC}$ locus, eight DNMs were observed in seven nsCL/P individuals and two DNMs in the NCR cohort (p = 0.02). Notably, the eight DNMs in nsCL/P clustered within 175 kb around the GWAS lead variant rs6740960. The enrichment at the 4q28.1 locus remained significant after correction for multiple testing for the number of TAD_{GWAS} (Figure 2D). No TAD_{GWAS} showed a significant depletion of nsCL/P DNMs. These results suggest at least two loci where both common and rare variants may contribute to nsCL/P risk, at $2p21_{PKDCC}$ presumably through regulatory effects on *PKDCC* (MIM: 614150).[41,42]

Identification of candidate TFs

Analyses were performed to test the hypothesis that DNMs contributing to nsCL/P might converge into

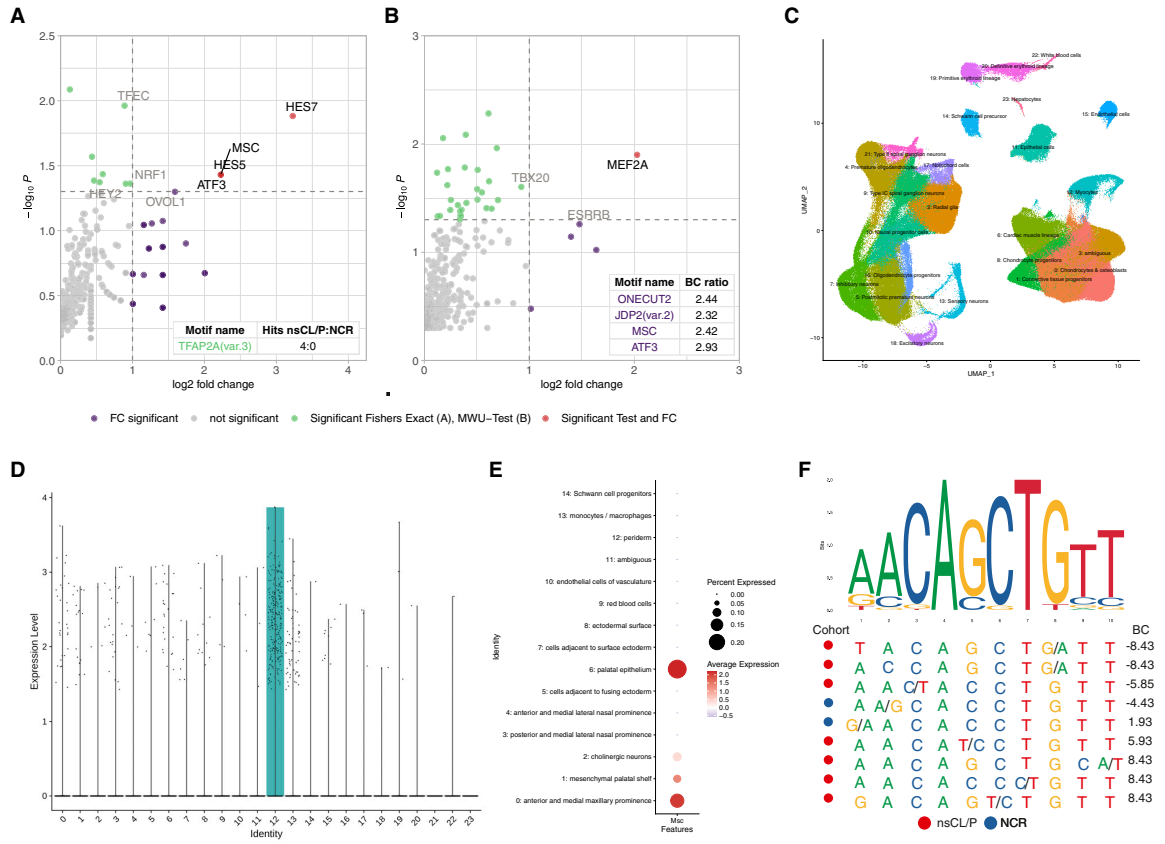


Figure 3. Identification of Musculin as a player in non-syndromic cleft lip with/without cleft palate etiology

(A) Qualitative analysis of DNMs in transcription factor (TF) binding sites (TFBS). Using 810 position weight matrices from JASPAR2020, the relative enrichment of non-syndromic cleft lip with/without cleft palate (nsCL/P) DNMs was assessed using log2FC (on y axis) versus Fisher's exact tests (−log10(p value) on x axis). Insert represents motif TFAP2a (var.3) that had log2FC ≥ 1 but lacked observations in the control cohort.

(B) Quantitative assessment of allelic effects on TF binding. For each DNM, the binding change (BC) of alternative versus reference allele was assessed via the Mann-Whitney U (MWU) test (on x axis) and log2FC (on y axis, calculated using the ratio of mean change of binding between cohorts). All motifs with ≥3 hits per cohort and sufficient variability in BCs were used for MWU testing. Inserts represent motifs that lacked sufficient observations for MWU testing, but had log2FC ≥ 1 and ≥5 hits.

(C–E) Single-cell transcriptomic data confirm a role for *Msc* during murine embryonic development.

(C) Re-analysis of MOCA data (Cao et al., 2019) identified 24 cell clusters at day E11.5.

(D) Expression levels for Musculin (*Msc*) in single-cell data from MOCA at E11.5 in cell clusters showed specific expression in myocytes (cell cluster 12 in C). Note: cluster numbers (x axis) correspond to cell cluster numbers in the UMAP plot in (C).

(E) Single-cell expression data of different cell clusters of the lambdoidal junction at E11.5 are shown as dot plot. For each cell cluster, the percentage of cells expressing *Msc* is indicated by dot size, while the average expression level is indicated by color. This illustrates expression of *Msc* in palatal epithelium and maxillary prominences.

(F) Nine DNMs mapped to the MSC motif (MA0665.1; seven in nsCL/P and two in NCR cohort). The sequences of the nine regions are illustrated per genomic region, as sorted according to BC, and with colored dots highlighting the cohort in which they were observed. At each position of a DNM, the allelic change is indicated in the order ref/alt.

molecular pathways through their location in transcription factor binding sites (TFBSs). Based on 28,773 DNMs and 810 PWMs, a total of 119,275 DNM-PWM hits were observed in the entire cohort. These pairs included 710 different PWMs and 21,043 DNMs (i.e., for 73.1% of the analyzed DNMs, the respective genomic context was located at a binding site of at least one PWM; Figure S11). After stringent filtering (supplemental methods), 88,129 DNM-PWM hits remained in the analysis. These showed a similar distribution in both cohorts (37,695 in nsCL/P versus 50,434 in NCR, $p = 0.56$).

At the level of individual PWMs, we observed four TFs whose PWMs showed a nominally significant excess in the nsCL/P trios (Figure 3A, HES7/HES5/ATF3/MSC; all $p < 0.05$), and a log2FC ≥ 1. In addition, 24 PWMs were identified for which at least one TFBS was predicted at a DNM region in the nsCL/P cohort, but none in the NCR cohort. These motifs included TFs with an established role in craniofacial development, such as TFAP2alpha (vers.3;

4 DNMs in nsCL/P, none in NCR; insert Figure 3A). When we aimed at identifying TF motifs with a significant difference in binding change (as opposed to frequency), one nominally significant hit (MEF2A, $p = 0.03$) was observed, together with an additional set of 17 motifs that had log2FC ≥ 1, but lacked the prerequisites for formal MWU calculations (supplemental methods; Figure 3B). Seven TFs were shared between the two approaches, including TFs Musculin (MSC; Table S27) and Activating Transcription Factor 3 (ATF3; Table S28). Notably, MSC and ATF3 were the only of these seven TFs for which a nominally significant Fisher's exact test result was generated (Table S29), prioritizing them as candidate TFs.

Analyses of single-cell expression data support a role for Musculin

Next, analyses were performed to determine the expression of the orthologs for *MSC* ([MIM: 603628]; *Msc*) and *ATF3* (*Atf3*) in single-cell data from the developing mouse embryo during E9.5 to E13.5 (MOCA[43]; Uniform Manifold Approximation and Projection [UMAP] plots in Figure S12). *Atf3* showed strong expression in endothelial cells, while being sparsely expressed in almost all other cell types (Figure S13). In contrast, our analyses revealed a specific expression pattern for *Msc* starting at E10.5. On day E10.5, *Msc* was expressed in sensory neurons but also in connective tissue progenitors and myocytes (Figure S14). Expression remained abundant in connective tissue progenitors, sensory neurons and myocytes on day E11.5 and was accompanied by expression in chondrocytes/osteoblasts and cardiac muscle lineage (Figures 3C and 3D). On day E12.5, *Msc* was most expressed in neural progenitor cells but also in sensory neurons and jaw and tooth progenitors. On day E13.5 *Msc* was expressed mainly in neural progenitor cells (Figure S14). While the MOCA data provide information on global expression in whole embryonic mice, their resolution concerning specific facial tissues is limited. Therefore, additional analyses were performed on single-cell data from the murine lambdoidal junction at day E11.5. Again, this revealed a low, but anatomically specific, expression of *Msc*, particularly in the palatal epithelium and the anterior and medial maxillary prominences (Figure 3E), while expression of *Atf3* was restricted to monocytes/macrophages and endothelial cells of vasculature (Figure S15).

DNMs in MSC binding sites affect binding *in vitro*

Based on those findings, we focused on MSC as candidate TF for nsCL/P. Detailed inspection of the MSC binding motifs revealed that the seven DNMs in nsCL/P were located at more central positions within the motifs, compared with the only two DNMs in the NCR cohort (Figure 3F; Table S27). To confirm that MSC binds to the predicted binding motif, and that binding is altered by the DNMs as predicted *in silico*, electrophoretic mobility shift assays (EMSAs) were performed for five selected DNMs, in triplicates.

For all five sequences, EMSA analysis confirmed the binding of MSC to either the reference and/or the alternative motif (Figure S16A; Table S30): for three of the five sequences, the observed direction of effect was consistent with predictions (i.e., gain of binding for chr. 16, loss of binding for chr. 5 and 10). For two regions, limited evidence was found for either any binding change at all (chr. 6), or the effect was observed in the opposite direction (chr. 7). Closer analysis of the respective genomic sequence revealed that, in the region of the DNM at chr. 7, a second MSC binding motif was present, which might have affected the prediction outcome (Figure S16B). The present data confirm that MSC binds to the predicted motif and suggest that this binding could be affected by mutations *in vitro*.

Discussion

WGS allows for a systematic investigation of genetic variants; i.e., across the allelic spectrum and variant types. Therefore, WGS data are a powerful resource to expand our understanding of susceptibility factors for nsCL/P, in particular when both coding and non-coding variants are analyzed jointly. However, the large number of rare variants in individual genomes challenges the identification of causal variants at the statistical level, and this is further hampered by our incomplete knowledge regarding regulatory processes occurring in the non-coding genome. In the present study, we analyzed DNMs as a specific class of variants, in a European-based nsCL/P cohort of 211 trios, and included both coding and non-coding variants in our investigation. While the cohort size is small compared with other traits of multifactorial etiology, it is similar to the cohort size included in the first nsCL/P GWAS that reported a genome-wide significant locus.[44] Three main findings emerged from our WGS study on nsCL/P.

First, while our study design included systematic approaches to enrich for true-positive signals, we failed to detect robust associations in our hypothesis-driven analyses. We observed some nominally significant findings, but these warrant further replication in order to allow for firm conclusions (in particular, for those findings that are based on singleton observations). Future studies including more trios and ethnicities but also additional control cohorts might be an important avenue to follow. The lack of systematic evidence in our study might indicate either that DNMs in the selected regions do not contribute to nsCL/P or that our analyses were statistically underpowered. Importantly, next to sample size, the power of our study might have been limited by the selection of the reference cohort, which comprised individuals with ES for which WGS data were generated within the same project. While this is a technical advantage for comparative analyses, some epidemiological data have suggested some shared etiology between OFC and cancer in general.[45] Still, so far, no evidence is available for a shared etiology between ES and nsCL/P from epidemiological or molecular data.[2] Furthermore, most current *in silico* prediction scores are trained on input data that are biased for deleterious

protein-coding variants and, therefore, are ineffective for non-coding regions. This limits their usage for WGS data, as illustrated in our study by the comparably low number of observed non-coding DNMs with high CADD scores.

Second, despite the limited evidence for overall enrichments, we identified a convergence of DNMs at loci that had prior evidence for an involvement in nsCL/P. Most interestingly, we observed a significant overrepresentation of DNMs in regions that were previously implicated in nsCL/P etiology by common variants. Specifically, two risk loci, 4q28 and 2p21$_{PKDCC}$, harbored significantly more DNMs in nsCL/P trios than the reference cohort. At 2p21, the variants clustered within a region of 175 kb, in close vicinity to rs6740960, which has been suggested as the sole causal variant at this locus.[39,46] As another example, we observed two intronic DNMs in the nsCL/P candidate gene, *ZFHX4*,[11] for which a frameshift mutation was previously reported (Table S31). While the exact functional effect and molecular mechanisms of these non-coding DNMs at GWAS loci or within candidate gene loci remain unclear, these findings illustrate the presence of allelic heterogeneity at established loci and pave the way for functional follow-up studies.

Finally, our results suggest that differential binding of Musculin (*MSC*, or MyoR) to its binding sequence might be of relevance to nsCL/P etiology. MSC is a basic-helix-loop-helix TF that is involved in the development of orofacial branchiomeric muscles (OBMs).[47] Interestingly, previous studies have identified sub-epithelial alterations in a specific OBM type, *musculus orbicularis oris*, as a subclinical phenotype in the relatives of individuals with nsCL/P, and these alterations are considered an intermediate phenotype of nsCL/P.[48–51] Notably, the network of TFs regulating OBM development includes several TFs that are encoded by genes implicated in nsCL/P via their presence at GWAS risk loci; i.e., *NOG* (MIM: 602991),[52] *PAX7* (MIM: 167410),[53] *FGF10* (MIM: 602115),[4] and *GREM1* (MIM: 603054)[54] (Figure S17). However, the exact coordination of this gene regulatory network and the context-specific effects of the binding changes remain unclear at the moment and require further investigation.

In summary, we here provide a genome-wide analysis of DNMs in nsCL/P that includes variation in the non-coding genome. While our study illustrates the challenges associated with our understanding of non-coding variation, we also provide evidence for causal DNMs at nsCL/P GWAS loci and suggest that common and rare variants in the muscle developmental pathway might be involved in nsCL/P etiology.

Data and code availability

Original data concerning the present genetic and functional analyses can be accessed as follows: WGS data for nsCL/P and NCR cohorts are available at dbGaP phs001168.v1.p1 and phs001228.v1.p1, respectively. Chromatin state segmentation data for craniofacial tissue (CT) are available at Gene Expression Omnibus (GEO), under accession number GSE97752. Chromatin state segmentation data for hNCC and cNCC are available at Zenodo (https://doi.org/10.5281/zenodo.3911187). CNEs are available on GitHub (https://github.com/pjshort/DDDNonCoding2017/tree/master/data). Original data of TADs are available at GEO under accession number GSE35156. Original data for single-cell expression from whole mouse embryos are available under https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads (Processed/Sampled/Split Data; gene_count_cleaned.RDS). Single-cell expression data for the lambdoidal junction are available at GEO under accession number GSM3867275. The accession number for the code of the modified version of denovoLOBGOB reported in this paper is publicly available at Zenodo (https://doi.org/10.5281/zenodo.5601707).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100166.

Acknowledgments

This work was supported by the German Research Council through funding provided to K.U.L. (DFG; LU 1944/3-1). H.K.Z. received support from the BONFOR program of the Medical Faculty Bonn (SciMed program, O-149.0132).

The present results were obtained using data generated by the Gabriella Miller Kids First (GMKF) Pediatric Research Program projects phs001168.v1.p1 and phs001228.v1.p1. Upon approved data access, data were downloaded from dbGaP (www.ncbi.nlm.nih.gov/gap) and the Website of the GMKF project (https//kidsfirstdrf.org). The GMKF Website and the Kids First Data Resource Center are supported by the National Institutes of Health (NIH) Common Fund (U2CHL138346). European nsCL/P trios were sequenced at Washington University's Mc Donnel Genome Institute (X01-HL132363, with principal investigators M.L.M. and E.F.) and this project was supported by the NIH through the following funding sources: R01-DE016148 (M.L.M. and S.M.W.), R01-DE014581 (T.H.B.), and R01-DD000295 (G.L.W.). Ewing sarcoma trios as NCR cohort were recruited within the context of the Children's Oncology Group AEPI10N5 Study (Genetic Epidemiology of Ewing Sarcoma, NCT01876303) and sequenced within the GMKF Ewing Sarcoma project (X01-HL132385, with principal investigator J.D.S.). The Ewing Sarcoma study was supported by the Children's Oncology Group and the National Cancer Institute.

Author contributions

H.K.Z. and K.U.L. conceptualized the study and acquired funding. H.K.Z., A. Schmidt, M.H., F.T., F.U.B., J.W., D.B., and P.M.K. analyzed sequencing data and/or provided computational resources. L.W. and H.K.Z. planned and performed statistical analyses. H.K.Z., L.W., A. Schmidt, A. Siewert, A.B.S., E.M., N.I., and K.U.L. jointly interpreted data. A. Siewert designed and performed the analysis of single-cell expression data. H.K.Z., S.A.J., and K.P. designed, performed, and interpreted EMSA experiments. H.K.Z. wrote the first version of the manuscript with contributions by L.W., A. Siewert, K.P., and K.U.L. All authors edited and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Received: June 1, 2022
Accepted: December 1, 2022

Web resources

GMKF Pediatric Research Program, www.commonfund.nih.gov/KidsFirst

denovoLOBGOB, https://github.com/pjshort/denovoTF.

FunciVar, https://github.com/Simon-Coetzee/funcivar.

GEO, https://www.ncbi.nlm.nih.gov/geo/

GENCODE, https://www.gencodegenes.org/human/grch37_mapped_releases.html.

GnomAD v3.1., https://gnomad.broadinstitute.org/

JASPAR 2020, https://bioconductor.org/packages/release/data/annotation/html/JASPAR2020.html.

MOCA, https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/landing.

OMIM, http://www.omim.org/.

TFBSTools, http://bioconductor.org/packages/release/bioc/html/TFBSTools.html.

Ensembl Variant Effect Predictor, https://www.ensembl.org/info/docs/tools/vep/online/input.html.

VISTA Enhancer Browser, https://enhancer.lbl.gov/

References

1. Mangold, E., Ludwig, K.U., and Nöthen, M.M. (2011). Breakthroughs in the genetics of orofacial clefting. Trends Mol. Med. *17*, 725–733.
2. Christensen, K., Juel, K., Herskind, A.M., and Murray, J.C. (2004). Long term follow up study of survival associated with cleft lip and palate at birth. BMJ *328*, 1405.
3. Grosen, D., Bille, C., Petersen, I., Skytthe, A., Hjelmborg, J.v.B., Pedersen, J.K., Murray, J.C., and Christensen, K. (2011). Risk of oral clefts in twins. Epidemiology *22*, 313–319.
4. Welzenbach, J., Hammond, N.L., Nikolić, M., Thieme, F., Ishorst, N., Leslie, E.J., Weinberg, S.M., Beaty, T.H., Marazita, M.L., Mangold, E., et al. (2021). Integrative approaches generate insights into the architecture of non-syndromic cleft lip ± cleft palate. HGG Adv. *2*, 100038.
5. Basha, M., Demeer, B., Revencu, N., Helaers, R., Theys, S., Bou Saba, S., Boute, O., Devauchelle, B., Francois, G., Bayet, B., et al. (2018). Whole exome sequencing identifies mutations in 10% of patients with familial non-syndromic cleft lip and/or palate in genes mutated in well-known syndromes. J. Med. Genet. *55*, 449–458.
6. Cox, L.L., Cox, T.C., Moreno Uribe, L.M., Zhu, Y., Richter, C.T., Nidey, N., Standley, J.M., Deng, M., Blue, E., Chong, J.X., et al. (2018). Mutations in the epithelial cadherin-p120-catenin complex cause mendelian non-syndromic cleft lip with or without cleft palate. Am. J. Hum. Genet. *102*, 1143–1157.
7. Savastano, C.P., Brito, L.A., Faria, Á.C., Setó-Salvia, N., Peskett, E., Musso, C.M., Alvizi, L., Ezquina, S.A.M., James, C., GOSgene, et al. (2017). Impact of rare variants in ARHGAP29 to the etiology of oral clefts: role of loss-of-function vs missense variants. Clin. Genet. *91*, 683–689.
8. Butali, A., Mossey, P., Adeyemo, W., Eshete, M., Gaines, L., Braimah, R., Aregbesola, B., Rigdon, J., Emeka, C., Olutayo, J., et al. (2014). Rare functional variants in genome-wide association identified candidate genes for nonsyndromic clefts in the African population. Am. J. Med. Genet. Part A *164A*, 2567–2571.
9. Letra, A., Maili, L., Mulliken, J.B., Buchanan, E., Blanton, S.H., and Hecht, J.T. (2014). Further evidence suggesting a role for variation in ARHGAP29 variants in nonsyndromic cleft lip/palate. Birth Defects Res. A Clin. Mol. Teratol. *100*, 679–685.
10. Leslie, E.J., Taub, M.A., Liu, H., Steinberg, K.M., Koboldt, D.C., Zhang, Q., Carlson, J.C., Hetmanski, J.B., Wang, H., Larson, D.E., et al. (2015). Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. Am. J. Hum. Genet. *96*, 397–411.
11. Bishop, M.R., Diaz Perez, K.K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J.B., Taub, M.A., Moreno-Uribe, L.M., Valencia-Ramirez, L.C., et al. (2020). Genome-wide enrichment of de novo coding mutations in orofacial cleft trios. Am. J. Hum. Genet. *107*, 124–136.
12. Fakhouri, W.D., Rahimov, F., Attanasio, C., Kouwenhoven, E.N., Ferreira De Lima, R.L., Felix, T.M., Nitschke, L., Huver, D., Barrons, J., Kousa, Y.A., et al. (2014). An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. Hum. Mol. Genet. *23*, 2711–2720.
13. Cvjetkovic, N., Maili, L., Weymouth, K.S., Hashmi, S.S., Mulliken, J.B., Topczewski, J., Letra, A., Yuan, Q., Blanton, S.H., Swindell, E.C., et al. (2015). Regulatory variant in FZD6 gene contributes to nonsyndromic cleft lip and palate in an African-American family. Mol. Genet. Genomic Med. *3*, 440–451.
14. Morris, V.E., Hashmi, S.S., Zhu, L., Maili, L., Urbina, C., Blackwell, S., Greives, M.R., Buchanan, E.P., Mulliken, J.B., Blanton, S.H., et al. (2020). Evidence for craniofacial enhancer variation underlying nonsyndromic cleft lip and palate. Hum. Genet. *139*, 1261–1272.
15. Shaffer, J.R., LeClair, J., Carlson, J.C., Feingold, E., Buxó, C.J., Christensen, K., Deleyiannis, F.W.B., Field, L.L., Hecht, J.T., Moreno, L., et al. (2019). Association of low-frequency genetic variants in regulatory regions with nonsyndromic orofacial clefts. Am. J. Med. Genet. Part A *179*, 467–474.
16. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
17. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. Am. J. Hum. Genet. *99*, 595–606.
18. Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.
19. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.
20. Huang, Y.F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.
21. Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding

pathogenic variants and putative essential regions of the human genome. Nat. Commun. *10*, 5241.
22. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. *17*, 122.
23. Jones, M.R., Peng, P.C., Coetzee, S.G., Tyrer, J., Reyes, A.L.P., Corona, R.I., Davis, B., Chen, S., Dezem, F., Seo, J.H., et al. (2020). Ovarian cancer risk variants are enriched in histotype-specific enhancers and disrupt transcription factor binding sites. Am. J. Hum. Genet. *107*, 622–635.
24. Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S.A., Swigut, T., and Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. Cell Stem Cell *11*, 633–648.
25. Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell *163*, 68–83.
26. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P., and Cotney, J. (2018). High-resolution epigenomic Atlas of human embryonic craniofacial development. Cell Rep. *23*, 1581–1597.
27. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., et al. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.
28. Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. *35*, D88–D92.
29. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92.
30. Li, H., Jones, K.L., Hooper, J.E., and Williams, T. (2019). The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. Development *146*, dev174888.
31. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.
32. Warner, D.R., Smith, H.S., Webb, C.L., Greene, R.M., and Pisano, M.M. (2009). Expression of Wnts in the developing murine secondary palate. Int. J. Dev. Biol. *53*, 1105–1112.
33. Geetha-Loganathan, P., Nimmagadda, S., Antoni, L., Fu, K., Whiting, C.J., Francis-West, P., and Richman, J.M. (2009). Expression of WNT signalling pathway genes during chicken craniofacial development. Dev. Dyn. *238*, 1150–1165.
34. Iyyanar, P.P.R., and Nazarali, A.J. (2017). Hoxa2 inhibits bone morphogenetic protein signaling during osteogenic differentiation of the palatal mesenchyme. Front. Physiol. *8*, 929.
35. Nie, S., Kee, Y., and Bronner-Fraser, M. (2009). Myosin-X is critical for migratory ability of Xenopus cranial neural crest cells. Dev. Biol. *335*, 132–142.
36. Hwang, Y.S., Luo, T., Xu, Y., and Sargent, T.D. (2009). Myosin-X is required for cranial neural crest cell migration in Xenopus laevis. Dev. Dyn. *238*, 2522–2529.
37. Bachg, A.C., Horsthemke, M., Skryabin, B.V., Klasen, T., Nagelmann, N., Faber, C., Woodham, E., Machesky, L.M., Bachg, S., Stange, R., et al. (2019). Phenotypic analysis of Myo10 knockout (Myo10tm2/tm2) mice lacking full-length (motorized) but not brain-specific headless myosin X. Sci. Rep. *9*, 597.
38. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
39. Ludwig, K.U., Böhmer, A.C., Bowes, J., Nikolić, M., Ishorst, N., Wyatt, N., Hammond, N.L., Gölz, L., Thieme, F., Barth, S., et al. (2017). Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. Hum. Mol. Genet. *26*, 829–842.
40. Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., Meng, L., Wang, W., Song, Y., Cheng, Y., et al. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat. Commun. *8*, 14364.
41. Imuta, Y., Nishioka, N., Kiyonari, H., and Sasaki, H. (2009). Short limbs, cleft palate, and delayed formation of flat proliferative chondrocytes in mice with targeted disruption of a putative protein kinase gene, Pkdcc (AW548124). Dev. Dyn. *238*, 210–222.
42. Melvin, V.S., Feng, W., Hernandez-Lagunas, L., Artinger, K.B., and Williams, T. (2013). A morpholino-based screen to identify novel genes involved in craniofacial morphogenesis. Dev. Dyn. *242*, 817–831.
43. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.
44. Birnbaum, S., Ludwig, K.U., Reutter, H., Herms, S., Steffens, M., Rubini, M., Baluardo, C., Ferrian, M., Almeida De Assis, N., Alblas, M.A., et al. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat. Genet. *41*, 473–477.
45. Bille, C., Winther, J.F., Bautz, A., Murray, J.C., Olsen, J., and Christensen, K. (2005). Cancer risk in persons with oral cleft - a population-based study of 8, 093 cases. Am. J. Epidemiol. *161*, 1047–1055.
46. Mohammed, J., Arora, N., Matthews, H.S., Hansen, K., Bader, M., Weinberg, S.M., Swigut, T., Claes, P., Selleri, L., Wysocka, J., et al. (2022). A common cis-regulatory variant impacts normal-range and disease-associated human facial shape through regulation of PKDCC during chondrogenesis. Preprint at bioRxiv. https://doi.org/10.1101/2022.09.05.506587.
47. Rosero Salazar, D.H., Carvajal Monroy, P.L., Wagener, F.A.D.T.G., and Von den Hoff, J.W. (2020). Orofacial muscles: embryonic development and regeneration after injury. J. Dent. Res. *99*, 125–132.
48. Weinberg, S.M., Neiswanger, K., Martin, R.A., Mooney, M.P., Kane, A.A., Wenger, S.L., Losee, J., Deleyiannis, F., Ma, L., De Salamanca, J.E., et al. (2006). The Pittsburgh Oral-Facial Cleft study: expanding the cleft phenotype. Background and justification. Cleft Palate. Craniofac. J. *43*, 7–20.
49. Martin, R.A., Hunter, V., Neufeld-Kaiser, W., Flodman, P., Spence, M.A., Furnas, D., and Martin, K.A. (2000). Ultrasonographic detection of orbicularis oris defects in first degree relatives of isolated cleft lip patients. Am. J. Med. Genet. *90*, 155–161.
50. Neiswanger, K., Weinberg, S.M., Rogers, C.R., Brandon, C.A., Cooper, M.E., Bardi, K.M., Deleyiannis, F.W.B., Resick, J.M., Bowen, A., Mooney, M.P., et al. (2007). Orbicularis oris muscle defects as an expanded phenotypic feature in nonsyndromic

cleft lip with or without cleft palate. Am. J. Med. Genet. Part A *143A*, 1143–1149.

51. Marazita, M.L. (2007). Subclinical features in non-syndromic cleft lip with or without cleft palate (CL/P): review of the evidence that subepithelial orbicularis oris muscle defects are part of an expanded phenotype for CL/P. Orthod. Craniofac. Res. *10*, 82–87.
52. Mangold, E., Ludwig, K.U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N.A., Chawa, T.A., Mattheisen, M., et al. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. Nat. Genet. *42*, 24–26.
53. Ludwig, K.U., Mangold, E., Herms, S., Nowak, S., Reutter, H., Paul, A., Becker, J., Herberz, R., AlChawa, T., Nasser, E., et al. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. Nat. Genet. *44*, 968–971.
54. Ludwig, K.U., Ahmed, S.T., Böhmer, A.C., Sangani, N.B., Varghese, S., Klamt, J., Schuenke, H., Gültepe, P., Hofmann, A., Rubini, M., et al. (2016). Meta-analysis reveals genome-wide significance at 15q13 for nonsyndromic clefting of both the lip and the palate, and functional analyses implicate GREM1 as a plausible causative gene. PLoS Genet. *12*, e1005914.

4.3 Combining genetic and single-cell expression data reveals cell types and novel candidate genes for orofacial clefting

Anna Siewert, Simone Hoeland, Elisabeth Mangold and Kerstin U. Ludwig

*Scientific Repo*rts 14:26492 (2024)

https://doi.org/10.1038/s41598-024-77724-9

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Combining genetic and single-cell expression data reveals cell types and novel candidate genes for orofacial clefting

Anna Siewert✉, Simone Hoeland, Elisabeth Mangold & Kerstin U. Ludwig✉

Non-syndromic cleft lip with/without cleft palate (nsCL/P) is one of the most common birth defects and has a multifactorial etiology. To date, over 45 loci harboring common risk variants have been identified. However, the effector genes at these loci, and the cell types that are affected by risk alleles, remain largely unknown. To address this, we combined genetic data from an nsCL/P genome-wide association study (GWAS) with single-cell RNA sequencing data obtained from the heads of unaffected human embryos. Using the recently developed single-cell disease relevance score (scDRS) approach, we identified two major cell types involved in nsCL/P development, namely the epithelium and the *HAND2*+ pharyngeal arches (PA). Combining scDRS with co-expression networks and differential gene expression analysis, we prioritized nsCL/P candidate genes, some of which were additionally supported by GWAS data (e.g., *CTNND1*, *PRTG*, *RPL35A*, *RAB11FIP1*, *KRT19*). Our results suggest that specific epithelial and PA sub-cell types are involved in nsCL/P development, and harbor a substantial fraction of the genetic risk for nsCL/P.

Keywords Cleft lip, Cleft palate, GWAS, scRNA-seq, hdWGCNA, Co-expression networks

Non-syndromic cleft lip with or without cleft palate (nsCL/P) is one of the most common birth defects, with a global prevalence of approximately 1 in 1,000 live births[1]. In addition to complex treatments such as surgery and speech therapy, affected patients are burdened by an increased risk of morbidity[2]. The etiology of nsCL/P involves both genetic and environmental factors[1]. To date, genome-wide association studies (GWAS) have identified more than 45 risk loci harboring common variants that are associated with increased nsCL/P risk[3]. At some loci, candidate genes have been pinpointed by evidence from syndromic forms of facial disorders, the presence of rare variants in affected individuals, or on the basis of results from animal models[4]. However, how the human risk variants affect the function of nsCL/P candidate genes and the cell types in which they are likely to act, remains largely unclear in most cases, although some initial work has been published[5–7].

To address these questions, single-cell RNA-sequencing (scRNA-seq) is a promising approach. Instead of analyzing gene expression profiles in bulk from whole tissues, scRNA-seq enables the investigation of gene expression profiles in specific individual cells. When applied to biomaterials of relevance to specific diseases, this allows both the generation of high-resolution transcription maps of cell-types, and the identification of sub-cell types that might contribute to disease pathogenesis.

In the context of craniofacial development, most scRNA-seq studies to date have been performed on murine tissues and have identified cell types that would have been missed in earlier analyses of bulk data. For example, one study identified heterogeneity in gene expression of mesenchymal cells in the anterior palate[8], while another found distinct cell populations at the fusion sites of the maxillary, medial-nasal, and lateral-nasal processes[9] in mice. To study the role of these cell types in nsCL/P, we and others have utilized these murine single cell expression maps to examine gene expression patterns of candidate genes identified in genetic studies[10,11]. However, the suitability of murine data for the investigation of nsCL/P is limited. Reasons for this include: (i) differences in morphology and tissue interactions between mice and humans, in particular during the later stages of facial development[12]; and (ii) the fact that in humans, most genetic nsCL/P associations are located in non-coding (and often non-conserved) regions of the genome, indicating higher-order regulatory mechanisms[4]. Recently, scRNA-seq data from unaffected human embryos aged four to six weeks were made available[13], which partly cover the crucial time period for nsCL/P development between the fourth and tenth week post-conception[14].

Institute of Human Genetics, School of Medicine & University Hospital Bonn, University of Bonn, Bonn, Germany. ✉email: anna.siewert@uni-bonn.de; kerstin.ludwig@uni-bonn.de

The joint study of genetic and transcriptomic data has the potential to identify affected cell types and improve the understanding of disease mechanisms during facial development. To date, few computational approaches that combine genetic and single-cell transcriptomic data have been available. However, the recently developed single-cell disease relevance score (scDRS)[15] now allows the identification of associations between candidate genes identified via GWAS and individual cells from scRNA-seq data. The aim of the present study was to identify human developmental cell types in which genetically-mediated nsCL/P risk is enriched, which is crucial in terms of unraveling the underlying molecular mechanisms of nsCL/P. For this purpose, the scDRS approach was used to combine scRNA-seq data of unaffected human embryos[13] with candidate genes derived from our recent GWAS on nsCL/P[3]. The identified cell types were then used to determine potential interactions between candidate genes in co-expression networks using high-dimensional weighted correlation network analysis (hdWGCNA)[16]. We demonstrate how these approaches can facilitate the identification of molecular networks, effector cell-types, and novel candidate genes, thus advancing our understanding of the molecular basis of genetic nsCL/P risk.

Methods

Human embryonic scRNA-seq data

Human embryonic scRNA-seq data[13] were downloaded from Gene Expression Omnibus (GSE157329, see data availability section). These data comprised scRNA-seq data from seven unaffected whole human embryos from Carnegie stage (CS) 12 (one embryo), CS 13–14 (three embryos), and CS 15–16 (three embryos), which had been broadly dissected into head, upper and lower trunk, limbs, and viscera. Based on meta-information provided by the authors, the data were reduced to dissection parts 'head' and 'head-upperTrunk'. This led to the exclusion of two embryos without head data, i.e., one embryo respectively from CS 13–14 and CS 15–16. No additional filtering was performed. These data were then re-analyzed using Seurat v4.3.0[17]. Details on analysis parameters are provided in the Supplementary Information. Briefly, to remove potential batch effects, data from different samples were integrated using canonical correlation analysis, as implemented in Seurat. For this purpose, the data were split according to sample ($n = 6$, including one donor head that was split and analyzed as two samples), and processed as individual Seurat objects prior to integration.

The data were normalized and 2,000 highly variable genes were identified before the data were scaled. For the integration of individual Seurat objects, integration anchors between objects were identified and then used to integrate the individual data sets into one data set. The resulting data set was scaled and cell cycle regression was performed as implemented in Seurat. Principal component analysis was performed using the variable features of the data. Clustering was performed by first identifying the shared nearest neighbors of cells and then clustering the cells using the original Louvain algorithm. The resulting data set contained 50,059 cells, which clustered into 25 cell clusters (between 276 and 4,993 cells per cluster, Fig. 1A). The clustering showed no influences attributable to sample batch effects (Fig. S1A). Cluster marker genes were determined (Table S1) and used for cell type annotation of the clusters, as based on the cell type marker genes from the original publication (Supplementary Table S1B from Xu et al. 2023). Identification of differentially expressed genes (DEGs) was performed for epithelial sub-clusters only (Table S2).

Identification of nsCL/P candidate cell types using scDRS

Preparation of the scRNA-seq data

To render the scRNA-seq data applicable for scDRS, the scaled data was removed and the Seurat object (containing only count and normalized data) was converted into an .h5ad file. Specific parameters are provided in the Supplementary Information.

Definition of nsCL/P gene set

For the preparation of the genetically-informed gene set, the candidate genes located in topologically associating domains (TADs) of nsCL/P GWAS risk loci were retrieved from Table S9 from Welzenbach et al. 2021[3]. Of the 404 genes located in these TADs, 51 were not detected in the scRNA-seq data. The remaining 353 genes are referred to as 'TAD genes'. Of these, 87 genes had reached significance in the gene-based test in the original publication, thus providing further genetic support for these genes beyond single-variation association statistics at the risk loci. These genes were used for scDRS analysis, and are referred to as 'nsCL/P gene set'. For all of these genes, z-scores were retrieved from the MAGMA output file from Welzenbach et al. 2021[3], and were used as weights in the "weighted setting".

Application of the scDRS method

Following download (GitHub) and installation of the scDRS package[15], scDRSs for all single cells in the scRNA-seq data ($n = 50,059$) were calculated. For the nsCL/P gene set, this was performed in both settings, i.e., "unweighted" and "weighted". Downstream analyses were conducted on the scores generated from each gene set. Specific analysis parameters are provided in the Supplementary Information.

Additional quality assessment

To identify potential artifacts, the entire scRNA-seq data set was used. Here, Pearson correlation between the scDRS of each cell and the total number of detected molecules of each cell was calculated.

Co-expression network analysis using hdWGCNA

To generate co-expression networks, the hdWGCNA package version 0.3.1 was used (see section Data availability). First, metacells were created from the single cell matrix, which were then normalized. Next, for both the epithelium and the *HAND2*+ PA, an expression matrix of the respective metacells was constructed. The soft

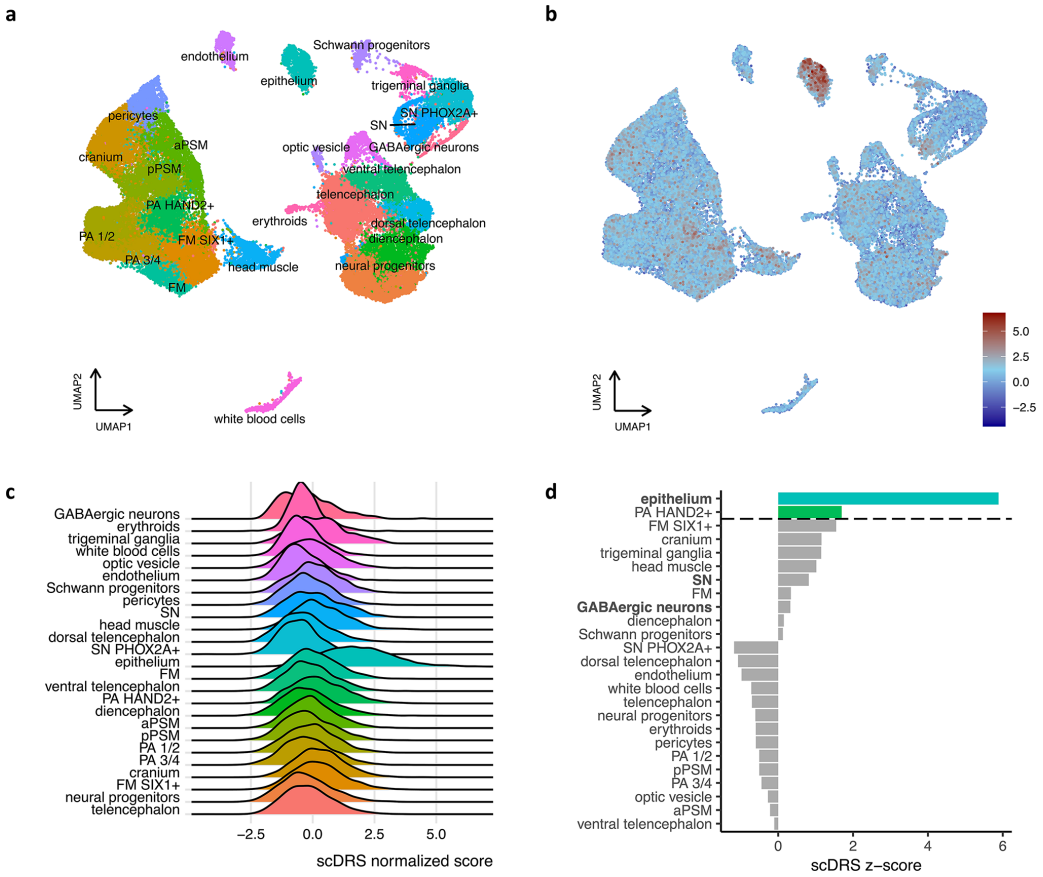


Fig. 1. scDRS identifies significant association with nsCL/P candidate genes in epithelium and *HAND2*+ pharyngeal arches (**a**) UMAP plot of scRNA-seq data from the heads of five unaffected human embryos from Carnegie stages 12–16. (**b**) UMAP plot from **a** colored according to the normalized scDRS for nsCL/P association at the single-cell level in the unweighted setting. (**c**) Ridgeplot of normalized scDRS according to cell type in the unweighted setting. (**d**) scDRS disease association at the cell type level in the unweighted setting. Cell types above the dashed line showed significant association with the nsCL/P gene set. Bold cell type labels indicate significant within-cell type heterogeneity in terms of disease association. Anterior presomitic mesoderm (aPSM), frontonasal mesenchyme (FM), $\log_2$ fold change ($\log_2$FC), pharyngeal arches (PA), posterior presomitic mesoderm (pPSM), sympathetic neurons (SN), single-cell disease relevance score (scDRS).

power thresholds were determined, the co-expression networks were constructed and the matrices were scaled. Then, module eigengenes and eigengene-based connectivity were determined. Specific analysis parameters are provided in the Supplementary Information. For each gene module, the percentage of TAD genes was calculated, and the three top gene modules, i.e., those with the highest percentage of TAD genes, were selected for further analysis. The Circos plots for the co-expression gene modules were created using the R package circlize[18].

Gene ontology enrichment analysis

A gene ontology (GO) analysis was performed using the clusterProfiler[19] and org.Hs.eg.db[20] R packages to identify biological process and molecular function GO terms. For E-9 and PA-14, redundant GO terms were removed (*simplify* function). Specific analysis parameters are provided in the Supplementary Information.

Identification of new candidate genes using hdWGCNA gene modules and GWAS data

For genes that were identified in the hdWGCNA gene modules, p-values from the gene-based test in our recent GWAS were retrieved from Table S6 from Welzenbach et al. 2021[3]. FDR correction was performed using the

Benjamini Hochberg method. For those genes that remained significant (adjusted p-value < 0.05) and were located outside of nsCL/P GWAS TADs, LocusZoom[21] was used to create regional association plots from the GWAS summary statistics[3]. To assess whether the co-expression gene modules were enriched for genes with a genetic association to nsCL/P, a gene set analysis was performed using MAGMA v1.10[22] (see section Data availability). The gene sets were created from the six selected co-expression modules (E-9, E-10, E11, PA-12, PA-14, PA-15). The analysis was performed using these gene sets and the MAGMA genes.raw output file from Welzenbach et al. 2021[3].

Results

Expression patterns of GWAS candidate genes implicate head epithelium and *HAND2*+ pharyngeal arches in genetically-mediated nsCL/P

Based on scRNA-seq data from the heads of unaffected human embryos (Fig. 1A) and the nsCL/P gene set informed by GWAS results (see Methods), developmental cell types that might underlie nsCL/P etiology were identified using scDRS.

First, scDRSs were calculated for each single cell in the scRNA-seq data ($n = 50{,}059$) in two settings, i.e., unweighted and weighted (by MAGMA z-scores, see Methods). At the single-cell level and across the two settings, an accumulation of cells with high scDRS was observed in the epithelium (Fig. 1B,C, Fig. S1B). When combining the scDRSs of individual cells over cell-clusters, this accumulation in the epithelium was found to be statistically significant in both the unweighted ($p = 0.002$, Fig. 1D, Table S3) and the weighted setting ($p = 0.01$, Fig. S1C, Table S4). In addition, the cell type *HAND2*+ PA reached statistical significance in the unweighted analysis ($p = 0.04$; Fig. 1D, Table S3). To identify potential subpopulations of disease-associated cells, all cell types were tested for evidence of within-cell type heterogeneity. In the unweighted setting, significant within-cell type heterogeneity was observed in the epithelium, as well as in sympathetic and GABAergic neurons (Fig. 1C). In the weighted setting, significant within-cell type heterogeneity was observed in the dorsal telencephalon, the sympathetic neurons, the endothelium, and the GABAergic neurons (Fig. S1C).To ensure that the scDRS for each cell and thereby the heterogeneity in association was not caused by technical differences in transcript detection, the correlation between the scDRS of each cell and the total number of molecules detected was tested. No strong support for such a technical bias was found (Pearson correlation coefficient: 0.006).

Given the converging evidence for a role in nsCL/P, and potential heterogeneity within the epithelial cell cluster, this cell cluster was then subdivided into two subclusters, as based on the scDRS p-value of each cell. This resulted in the identification of an associated subcluster (435 cells, $p \leq 0.01$) and a non-associated subcluster (717 cells, $p \geq 0.1$), from a total of 1,835 cells. DEGs between these subclusters were then identified. A total of 139 DEGs showed higher expression in the subcluster of disease-associated cells (fold change > 1) compared to the subcluster containing non-associated cells (fold change < -1; Fig. 2; Table S2). Of these, 31 genes were among the 353 'TAD genes', and 25 of these 31 genes were among the 87 genes of the 'nsCL/P gene set' of

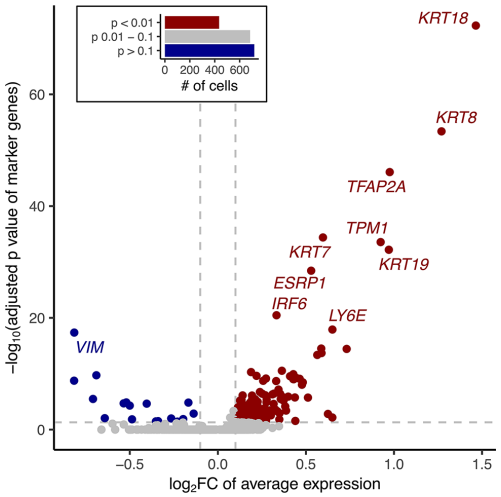


Fig. 2. Marker genes of associated epithelial cells contain known nsCL/P candidate genes. Volcano plot showing differentially expressed genes between nsCL/P associated (scDRS p-value < 0.01) and non-associated (scDRS p-value > 0.1) epithelial cells. Numbers within each group are depicted in the integrated bar plot. Genes with adjusted p-values < 0.05 (dashed horizontal line) and $\log_2$FC > 0.1 were considered marker genes for nsCL/P associated cells (red). Genes with adjusted p-values < 0.05 and $\log_2$FC < -0.1 were considered marker genes for non-associated cells (blue). Top 10 genes with the lowest p-values are labeled. Non-syndromic cleft lip with/without cleft palate (nsCL/P), $\log_2$ fold change ($\log_2$ FC), single-cell disease relevance score (scDRS).

previously suggested effector genes (e.g., *KRT8*, *KRT18*, *TFAP2A*, *TPM1*, *ESRP1*, and IRF6; Fig. 2; Table S2). Based on these results, we prioritized the remaining six TAD genes as nsCL/P candidate genes (Table 1, 'DEG' approach). Notably, for some of them, evidence of an involvement in orofacial clefting phenotypes has already been presented, though not from GWAS data[23–25].

scDRS-informed prioritization of candidate genes at GWAS loci

In addition to confirming known nsCL/P risk genes, scDRS within associated cell types can also be used to identify potential novel candidate genes. For this purpose, we used the gene-level downstream application of scDRS, which correlates the scDRS with the expression of genes that are not part of the tested gene set. Specifically, we aimed to identify genes at GWAS loci that were not prioritized as candidate genes due to the presence of another promising gene, or due to the lack of a significant gene-based P-value in the genetic data[3]. In this analysis, we considered genes with a correlation coefficient of > 0.01. In the unweighted setting, a positive correlation was observed between gene expression and scDRS for 33 genes (Table 1 'scDRS TAD gene' approach, Table S5), with the highest absolute value being observed for *RPL35A* (Pearson correlation coefficient: 0.15). This gene is located at the 3q29 locus, which harbors the previously proposed candidate genes *DLG1* and *MELTF*. The TAD genes *GADD45B* (19p13.3), *ARHGAP29* (1p22), and *MSX1* (4p16.2) showed a positive, albeit less pronounced correlation (Pearson correlation coefficients between 0.048 and 0.055). Our findings provide additional support for prioritizing these as effector genes at their respective loci. Similar results were found in the weighted setting (Table S6).

Co-expression network analysis of epithelium and *HAND2*+ pharyngeal arches

To identify genes with potential interactions in the previously identified nsCL/P-associated cell types (epithelium and *HAND2*+ PA), co-expression networks were generated. For each cell type, hdWGCNA identified 18 groups of interconnected, positively correlated genes (so-called 'gene modules'). Of these, three per cell type were selected for further analysis (see Methods, Tables S7 & S8).

Of the epithelial gene modules, the following were selected: E-9 (348 genes / including 12 nsCL/P TAD genes); E-10 (73/3); and E-11 (201/8) (Fig. 3A, Table S7). The eigengene values within the epithelial cluster were then plotted in their UMAP space of scRNA-seq data (Fig. 3C). While for E-9 and E-11, these appeared to be restricted to the upper part and lower part of the UMAP plot respectively, the highest values for E-10 did not appear specific. For E-9, the hub genes (i.e., genes with the largest number of connections within the module's network) included the nsCL/P TAD genes *TFAP2A*, *TPM1*, and *ARHGAP29*, thus providing further support for the hypothesis that they play a causal role in nsCL/P at their respective loci (Table S9)[24,26]. Enrichment analyses for E-9 using GO terms identified odontogenesis, wound healing, actin filament organization, the canonical Wnt signaling pathway, and Cadherin binding (Fig. 3D, Table S10). In contrast to E-9, the hub genes of E-10 and E-11 contained no nsCL/P TAD genes, and the GO term analysis results were non-specific (pituitary gland development and central nervous system neuron differentiation for E-10; regulation of neuron differentiation, ribosome binding, and unfolded protein binding for E-11; Table S10). Together, this suggests that the most relevant epithelial gene module in terms of nsCL/P risk may be E-9.

For *HAND2*+ PA, the selected gene modules were: PA-12 (111 genes / including 5 nsCL/P TAD genes), PA-14 (145/7), and PA-15 (83/4; Fig. 3B, Table S8). The highest module eigengene values for PA-12 and PA-15 were distributed evenly over the cluster, while the highest values for PA-14 were more concentrated at the bottom of the UMAP space (Fig. 3C). For PA-12, the identified hub genes included *MRC2*, which is located at the nsCL/P risk locus 17q23.2 but showed no significant gene-based association in Welzenbach et al. 2021[3]. Notably, while a set of seven genes is located at this locus, none has garnered sufficient research evidence to date to be considered the effector gene. Therefore, our results now prioritize *MRC2* as a candidate gene for functional studies. The hub genes of PA-14 contained the nsCL/P candidate genes *TPM1* and *ZFP36L2*, while those in PA-15 included the nsCL/P candidate gene *KRT18* (Table S11). The GO term analysis generated no significantly enriched terms for PA-12 or PA-15 (Table S10). However, for PA-14, the GO terms included T-cell differentiation and the development of the roof of the mouth, tongue, muscle tissue, arteries, and mesenchyme (Fig. 3D, Table S10).

When comparing the three epithelial gene modules and the three *HAND2*+ PA gene modules, a limited overlap of between 1 and 11 genes was observed (Fig. 3E, Fig. S2A/B, Table S12 e.g. *ZFP36L2* and *TPM1*). This suggests that most genes located at GWAS loci act or interact in only one of the two cell types.

Analysis approach	Prioritized nsCL/P candidate genes
scDRS TAD gene	***RPL35A***, ***C10orf82***, *MSX1*, ***GADD45B***, *ARHGAP29*, ***TLE2***, ***NSD3***, ***TIMM13***, ***RAB11FIP1***, ***NBL1***
scDRS gene	***KRT19***, ***CLDN6***, ***EPCAM***, ***CLDN4***, ***CLDN7***, ***RAB25***, ***RPS14***, ***RPL41***, ***RPS18***, ***AP1M2***
hdWGCNA + GWAS	*CTNND1*, *PRTG*, ***BFAR***, ***HYAL2***
DEGs	*ARHGAP29*, *MYC*, ***GADD45B***, ***RAB11FIP1***, ***PLKHF2***, ***NSG1***

Table 1. Summary of potential new candidate genes for nsCL/P based on different analysis approaches. The top 10 genes per analysis approach are shown, the remaining genes for scDRS TAD gene and scDRS gene are listed in table S5. Genes that have not been previously reported with nsCL/P are shown in bold. Analysis approaches are defined in the main text. Abbreviations: Differentially expressed genes (DEGs), genome-wide association study (GWAS), high-dimensional weighted correlation network analysis (hdWGCNA), non-syndromic cleft lip with or without cleft palate (nsCL/P), single-cell disease relevance score (scDRS), topologically associating domain (TAD).

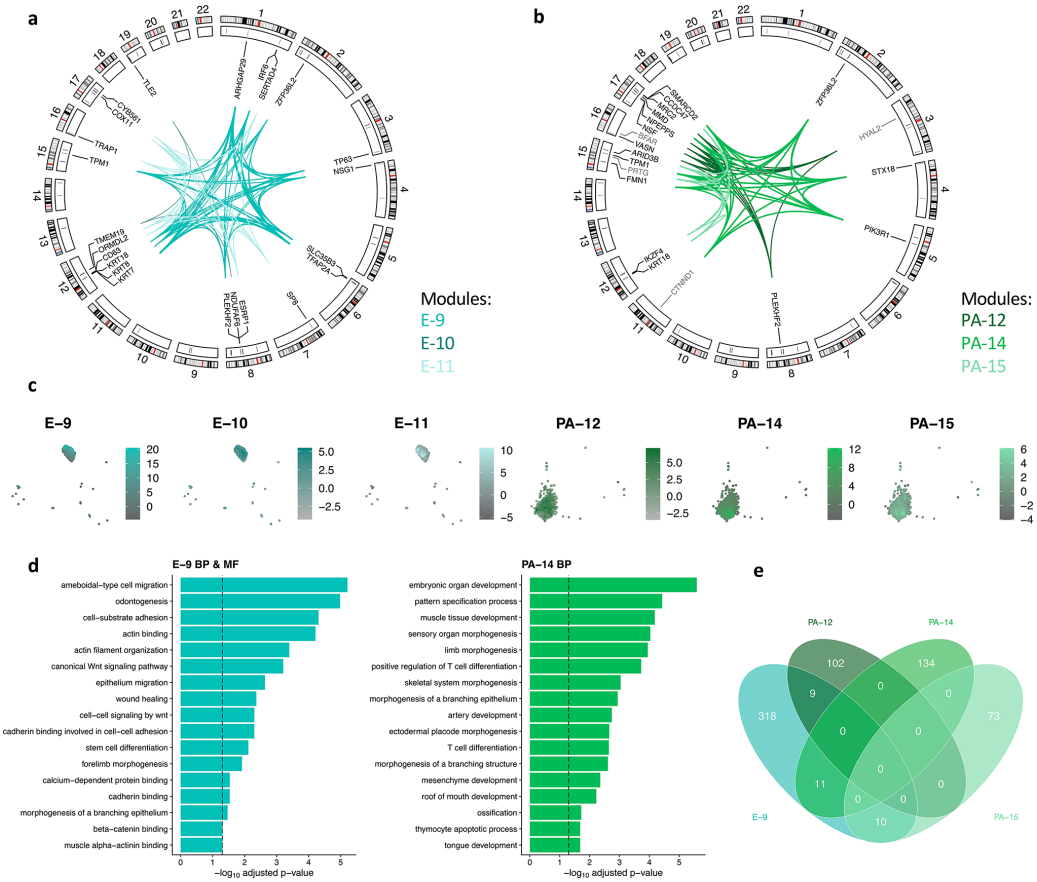


Fig. 3. Co-expression gene modules of epithelium and *HAND2*+ PA. (**a**) Circos plot of nsCL/P genes in epithelial co-expression gene modules. The outer track shows the chromosomal cytoband, the inner track shows the positions of TADs described in Welzenbach et al. 2021[3]. The colors of the connecting lines correspond to the respective gene module. The strength of the connecting lines reflects the pairwise correlation coefficient between two genes multiplied by a factor of 30 for illustration purposes. (**b**) Circos plot of nsCL/P genes (black) and potential novel candidate genes (gray) in *HAND2*+ PA co-expression gene modules. Panel layout as described in **a**. (**c**) UMAP plots of epithelium (E-9, E-10, E-11) and *HAND2*+ PA (PA-12, PA-14, PA-15) colored according to the module eigengene values for each co-expression gene module. (**d**) Bar plots of selected GO terms for biological process (E-9 & PA-14) and molecular function (E-9) for the epithelial co-expression gene module E-9 and the *HAND2*+ PA co-expression gene module PA-14. The vertical dashed line is set at p-value -$\log_{10}$ of 0.05. (**e**) Venn diagram of gene overlap between epithelial gene module E-9 and *HAND2*+ PA gene modules PA-12, PA-14, and PA-15. Epithelium (E), gene ontology (GO), non-syndromic cleft lip with/without cleft palate (nsCL/P), pharyngeal arches (PA), topologically associating domains (TADs).

MAGMA gene set analysis identifies association between epithelial co-expression gene module and nsCL/P

To examine the joint association of genes within hdWGCNA-identified co-expression gene modules and nsCL/P, a MAGMA gene set analysis was performed using the six selected gene modules as individual gene sets (see above) and the nsCL/P GWAS summary statistics from Welzenbach et al. 2021[3]. This analysis revealed a significant association with nsCL/P for gene module E-9 ($p = 5.98 \times 10^{-5}$), and provides further evidence that this gene module is enriched with genes that are associated with nsCL/P (Table S13).

Identification of novel nsCL/P candidate genes

To identify potential novel nsCL/P candidate genes, an analysis was performed of genes that are located outside the GWAS-TADs and which represent plausible novel candidate genes given similarities in scDRS and co-expression networks. First, genes whose expression patterns are positively correlated (correlation coefficient > 0.01) with the scDRS of individual cells (see above), but which are located outside of any known GWAS locus, were identified. Here, *KRT19*, *EPCAM*, and the Claudin family members *CLDN6*, *CLDN4*, and *CLDN7* were moderately correlated with the scDRS in the unweighted setting (Pearson correlation coefficients 0.17 to 0.29; Table 1 'scDRS gene' approach, Table S5). None of these genes showed significance in the MAGMA gene-based test in Welzenbach et al. 2021[3]. However, *KRT19* and *EPCAM* were among the epithelial co-expression modules (E-9 and E-11, respectively). Similar results were obtained for the weighted setting (Table S6).

Second, genes that were listed among the selected hdWGCNA gene modules (E-9, E-10, E-11, PA-12, PA-14, and PA-15), and which were significant in the gene-based test, were examined. No new candidate genes were identified in any of the three epithelial gene modules. However, evidence was generated to suggest that *HYAL2* and *BFAR* (both in PA-12), *CTNND1* (PA-14), and *PRTG* (PA-15) represent novel nsCL/P candidate genes from the *HAND2*+ PA gene modules (Table 1 "hdWGCNA + GWAS' approach, Table S8). An examination of the association structure around these genes in the GWAS summary statistics yielded nominally significant genetic support for the loci harboring *CTNND1* and *PRTG* (Fig S2 C/D). This suggests that these loci might reach conservative thresholds for genetic associations of common variation in future studies involving increased power. Interestingly, previous studies already linked rare variants in *CTNND1*[27] and low-frequency coding variants in *PRTG*[28] to nsCL/P.

Discussion

In recent years, multiple genetic studies on nsCL/P have identified genomic risk loci, and suggested local candidate genes in the associated regions[3,25,29–43]. However, since most of the associated regions map to non-coding parts of the genome and can thus be presumed to have context-specific effects, biological interpretation of these discoveries requires the identification of the affected cell types. This knowledge would in turn inform the context in which functional studies should be performed, which is essential for understanding the molecular mechanisms of nsCL/P development. For some established candidate genes, expression patterns have already been reported, e.g., *IRF6* expression in neural crest and epithelial cells[6,44–48], and *TFAP2A* expression in facial mesenchyme, nervous system, epithelial, and neural crest cells[49–51]. In addition, in a previous study involving single-cell transcriptome analyses in mice[10], our group showed that the murine homologs of certain nsCL/P candidate genes are expressed predominantly in either epithelial cell types (e.g., candidate genes *IRF6*, *TFAP2A*, *ESRP1*) or mesenchymal-like cell types (e.g., candidate genes *ALX1*, *ALX3*, *GREM1)*. The present study complemented previous research by performing a systematic examination of the joint gene expression of nsCL/P candidate genes from GWAS, with the aim of detecting the human developmental cell types that mediate genetic nsCL/P risk.

Based on expression data from unaffected human embryonic heads, our scDRS analysis implicated the epithelium and *HAND2*+ PA as primary cell types with an involvement in genetic nsCL/P risk. This confirms, and further refines, observations from our previous study in mice, which showed that individual nsCL/P genes were expressed in epithelial and mesenchymal cell types[10]. Epithelial cells are involved in manifold processes during lip and palate formation. These processes include: (i) epithelial seam formation, which is required for the fusion processes of the upper lip and the palate, as well as those between the medial-nasal, the lateral nasal, and the maxillary prominences[52–55]; (ii) epithelial-to-mesenchymal transitions, which allow for movement of cells as facial prominences grow, as well as the removal of epithelial seams[56–58]; (iii) the formation of the periderm, which covers the developing epithelium[59]; and (iv) cell adhesion and migration[60,61]. We speculate that the heterogeneity we observed in the overall epithelial cell cluster might recapitulate different expression patterns associated with these different functions. Indeed, one of the top markers of the associated epithelial cells is *IRF6*, which is a particularly relevant gene in periderm formation, and has been shown to be crucial in the development of the palate[62,63]. However, we note that the data include the whole head and, therefore, the possibility remains that the within-cell type heterogeneity might also be caused by epithelial cells originating from other regions of the head, rather than the facial processes. The second major cell-type we identified were *HAND2*+ PA, one of three clusters that were annotated as PA, which give rise to the bones and connective tissue of the head[64]. The transcription factor *HAND2*, which characterized the specific PA cluster associated with nsCL/P, was previously found to be expressed in the neural-crest derived mesenchyme of the PA of mice[65] and involved with patterning in the PA of zebrafish[66]. Together, the findings of specific cell types within PA and epithelium suggest that there may be specific sub-cell types that are involved in nsCL/P, which should be assessed with spatial and functional data in future studies.

Having identified relevant cell-clusters, the respective expression data can be explored using co-expression network analysis in order to identify genes that are potentially subject to the same gene regulation or which interact on a molecular level. This can help to prioritize effector genes at established genomic risk loci or identify new candidate genes. Importantly, our co-expression modules identified known interactions, such as *IRF6* and *TFAP2A*, which have been shown to act jointly in a genetic pathway[5,6], as well as *IRF6* and *TP63*, the latter of which has been reported to activate *IRF6* expression[67]. This suggests that some of the newly identified genes within the same gene modules are promising genes for further functional studies, for example, *TPM1* and *ZFP36L2*, which occurred in the same gene module in the epithelial cells and in the *HAND2*+ PA. Interestingly, recent studies found that *ZFP36L2* was significantly associated with nsCL/P and one of its subtypes, i.e., non-syndromic cleft lip only, in GWAS data from a Chinese Han population[68,69]. We also found evidence for a role of *RPL35A*, which is located in a larger deletion region in patients with craniofacial abnormalities[70]. Our data provide further support for the suspected genes *GADD45B*, *ARHGAP29*, and *MSX1*, though these had not been

prioritized in Welzenbach et al. 2021[3]. The present analyses also identified new candidate genes located outside of GWAS loci, such as *CTNND1* and *PRTG*, which were implicated through their expression patterns in the *HAND2*+ PA gene modules. For both genes, genetic support is provided by our in-house GWAS data, but also through rare variants identified by exome sequencing of multiplex families for *CTNND1*[27] and low-frequency coding variants in *PRTG*[28].

To obtain information on the biological relevance of the gene modules, we performed GO term analyses. The results reflect the before discussed processes the epithelium and PA are involved in, e.g. epithelial morphogenesis, mesenchyme development, and migration. Additionally, they support functional hypotheses such as cadherin-binding via *CTNND1*, involvement of several members of the Wnt-family, and muscle tissue development, all of which have been previously implicated in nsCL/P[27,71–78]. The association of one gene module with T-cell differentiation could provide an exciting link to immunological factors, which requires further examination. The differences in biological functions between the gene modules of epithelium and PA together with a very small gene overlap in genes between the gene modules, suggest that the genetic risk for nsCL/P is split on different biological, and maybe complementary, functions across those two cell types.

In summary, we combined human scRNA-seq data with genetic information on nsCL/P risk and identified nsCL/P-associated cell types and potential sub-cell types, which might harbor a considerable part of the genetic risk. Co-expression networks in these cell types allowed us to identify established and potential new gene-gene interactions. We also demonstrated how to identify new candidate genes based on these networks by revisiting the initial GWAS data.

Data availability

The original scRNA-seq data from Xu et al. are available via Gene Expression Omnibus accession number GSE157329 or via https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157329. Our re-analyzed scRNA-seq data have been deposited at Zenodo in Seurat object format (DOI: 10.5281/zenodo.12742819). hdWGCNA documentation: https://smorabit.github.io/hdWGCNA/index.html. MAGMA: https://cncr.nl/research/magma/.

Received: 22 August 2024; Accepted: 24 October 2024
Published online: 03 November 2024

References

1. Mangold, E., Ludwig, K. U. & Nöthen, M. M. Breakthroughs in the genetics of orofacial clefting. *Trends Mol. Med.* **17** (12), 725–733 (2011).
2. Christensen, K., Juel, K., Herskind, A. M. & Murray, J. C. Long term follow up study of survival associated with cleft lip and palate at birth. *Br. Med. J.* **328** (7453), 1405–1406 (2004).
3. Welzenbach, J. et al. Integrative approaches generate insights into the architecture of non-syndromic cleft lip with or without cleft palate. *Hum. Genet. Genomics Adv.* **2** (3), 1–14 (2021).
4. Thieme, F. & Ludwig, K. U. The role of noncoding genetic variation in isolated Orofacial Clefts. *J. Dent. Res.* **96** (11), 1238–1247 (2017).
5. Carroll, S. H. et al. An Irf6-Esrp1/2 regulatory axis controls midface morphogenesis in vertebrates. *Dev*, **147** (24) (2020).
6. Kousa, Y. A. et al. The TFAP2A-IRF6-GRHL3 genetic pathway is conserved in neurulation. *Hum. Mol. Genet.* **28** (10), 1726–1737 (2019).
7. Kousa, Y. A., Fuller, E. & Schutte, B. C. IRF6 and AP2A Interaction regulates epidermal development. *J. Invest. Dermatol.* **138** (12), 2578–2588 (2018).
8. Ozekin, Y. H., O'Rourke, R. & Bates, E. A. Single cell sequencing of the mouse anterior palate reveals mesenchymal heterogeneity. *Dev. Dyn.* **252** (6), 713–727 (2023).
9. 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. *Dev.* **146** (12) (2019).
10. Siewert, A. et al. Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data. *Front. Cell Dev. Biol.*, **11** (April), 1–11 (2023).
11. Cui, X. et al. Genetic variants in BCL-2 family genes influence the risk of non-syndromic cleft lip with or without cleft palate. *Birth Defects Res.* **116** (1), 1–12 (2024).
12. Yu, K., Deng, M., Naluai-Cecchini, T., Glass, I. A. & Cox, T. C. Differences in oral structure and tissue interactions during mouse vs. human palatogenesis: Implications for the translation of findings from mice. *Front. Physiol.* **8** (Mar), 1–12 (2017).
13. Xu, Y. et al. A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nat. Cell. Biol.* 1–12 (2023).
14. 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** (3), 167–178 (2011).
15. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* (2022).
16. Morabito, S., Reese, F., Rahimzadeh, N. & Miyoshi, E. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell. Rep. Methods*. **3** (6), 100498 (2023).
17. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell*. **184** (13), 3573–3587.e29, (2021).
18. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics*. **30** (19), 2811–2812 (2014).
19. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*. **2** (3), 100141 (2021).
20. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.8.2. (2019).
21. Boughton, A. P. et al. LocusZoom.js: Interactive and embeddable visualization of genetic association study results. *Bioinformatics*. **37** (18), 3017–3018 (2021).
22. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11** (4), 1–19 (2015).
23. Leslie, E. J. et al. Expression and mutation analyses implicate ARHGAP29 as the etiologic gene for the cleft lip with or without cleft palate locus identified by genome-wide association on chromosome 1p22. *Birth Defects Res. Part. - Clin. Mol. Teratol.* **94** (11), 934–942 (2012).
24. Savastano, C. P. et al. Impact of rare variants in ARHGAP29 to the etiology of oral clefts: Role of loss-of-function vs missense variants. *Clin. Genet.* **91** (5), 683–689 (2017).

25. Yu, Y. et al. Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat. Commun.* **8** (2), 1–11 (2017).
26. Milunsky, J. M. et al. TFAP2A mutations result in branchio-oculo-facial syndrome. *Am. J. Hum. Genet.* **82** (5), 1171–1177 (2008).
27. Cox, L. L. et al. Mutations in the epithelial Cadherin-p120-Catenin complex cause mendelian non-syndromic cleft lip with or without cleft palate. *Am. J. Hum. Genet.* **102** (6), 1143–1157 (2018).
28. Leslie, E. J. et al. Association studies of low-frequency coding variants in nonsyndromic cleft lip with or without cleft palate. *Am. J. Med. Genet. Part. A.* **173** (6), 1531–1538 (2017).
29. Rahimov, F. et al. Disruption of an AP-2α binding site in an IRF6 enhancer is strongly associated with cleft lip. *Nat. Genet.* **40** (11), 1341–1347 (2008).
30. Ludwig, K. U. et al. Meta-analysis reveals genome-wide significance at 15q13 for nonsyndromic clefting of both the lip and the Palate, and functional analyses implicate GREM1 as a plausible causative gene. *PLoS Genet.* **12** (3), 1–21 (2016).
31. Leslie, E. J. et al. Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Hum. Genet.* **136** (3), 275–286 (2017).
32. Ludwig, K. U. et al. Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. *Hum. Mol. Genet.* **26** (4), 829–842 (2017).
33. Mostowska, A. et al. Common variants in DLG1 locus are associated with non-syndromic cleft lip with or without cleft palate. *Clin. Genet.* **93** (4), 784–793 (2018).
34. 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** (2), 215–226 (2020).
35. Mukhopadhyay, N. et al. Genome-wide association study of multiethnic nonsyndromic orofacial cleft families identifies novel loci specific to family and phenotypic subtypes. *Genet. Epidemiol.* **46**, 3–4 (2022).
36. Moreno, L. M. et al. FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Hum. Mol. Genet.* **18** (24), 4879–4896 (2009).
37. Birnbaum, S. et al. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat. Genet.* **41** (4), 473–477 (2009).
38. Beaty, T. H. et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat. Genet.* **42** (6), 525–529 (2010).
39. Mangold, E. et al. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat. Genet.* **42** (1), 24–26 (2010).
40. Ludwig, K. U. et al. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci, **44** (9), 968–971 (2012).
41. Beaty, T. H. et al. Confirming genes influencing risk to cleft lip with/without cleft palate in a case-parent trio study. *Hum. Genet.* **132** (7), 771–781 (2013).
42. Sun, Y. et al. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat. Commun.*, **6** (2015).
43. Leslie, E. J. et al. A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p 24.2, 17q23 and 19q13. *Hum. Mol. Genet.* **25** (13), 2862–2872 (2016).
44. Kondo, S. et al. Mutations in IRF6 cause Van Der Woude and popliteal pterygium syndromes. *Nat. Genet.* **32** (2), 285–289 (2002).
45. Knight, A. S., Schutte, B. C., Jiang, R. & Dixon, M. J. Developmental expression analysis of the mouse and chick orthologues of IRF6: The gene mutated in Van Der Woude syndrome. *Dev. Dyn.* **235** (5), 1441–1447 (2006).
46. Richardson, R. J., Dixon, J., Jiang, R. & Dixon, M. J. Integration of IRF6 and Jagged2 signalling is essential for controlling palatal adhesion and fusion competence. *Hum. Mol. Genet.* **18** (14), 2632–2642 (2009).
47. Goudy, S. et al. Cell-Autonomous and non-cell-autonomous roles for Irf6 during development of the Tongue. *PLoS One*, **8** 2 (2013).
48. Kousa, Y. A. & Schutte, B. C. Toward an orofacial gene regulatory network. *Dev. Dyn.* **245** (3), 220–232 (2016).
49. Schorle, H., Meiert, P., Buchertt, M., Jaenisch, R. & Mitchellt, P. J. Transcription factor AP-2 essential for cranial closure and craniofacial development. **381** (May), 235–238 (1996).
50. Nguyen, T. T. et al. TFAP2 paralogs regulate midfacial development in part through a conserved ALX genetic pathway. *Dev*, **151** (1) (2024).
51. Woodruff, E. D., Gutierrez, G. C., Van Otterloo, E., Williams, T. & Cohn, M. J. Anomalous incisor morphology indicates tissue-specific roles for Tfap2a and Tfap2b in tooth development. *Dev. Biol.*, **472** (January) 67–74 (2021).
52. Gaare, J. D. & Langman, J. Fusion of nasal swellings in the mouse embryo: Regression of the nasal fin. *Am. J. of Anatomy*. **150**, 477–499 (1977).
53. Gaare, J. D. & Langman, J. Fusion of nasal swellings in the mouse embryo: Surface coat and initial contact. *Am. J. of Anatomy*. **1503** (3), 461–475 (1977).
54. Abramyan, J. & Richman, J. M. Recent insights into the morphological diversity in the amniote primary and secondary palates. *Dev. Dyn.* **244** (12), 1457–1468 (2015).
55. Kim, S. et al. Convergence and extrusion are required for Normal Fusion of the mammalian secondary palate. *PLoS Biol.* **13** (4), 1–24 (2015).
56. Losa, M. et al. Face morphogenesis is promoted by pbx-dependent EMT via regulation of snail1 during frontonasal prominence fusion. *Dev*. **145** (5) (2018).
57. Fitchett, J. E. & Hay, E. D. Medial edge epithelium transforms to mesenchyme after embryonic palatal shelves fuse. *Dev. Biol.* **131** (2), 455–474 (1989).
58. Jin, J. Z. & Ding, J. Analysis of cell migration, transdifferentiation and apoptosis during mouse secondary palate fusion. *Development*. **133** (17), 3341–3347 (2006).
59. Merker, H. & V. M'Boneko and Development and morphology of the periderm of mouse embryos (days 9–12 of gestation). *Acta Anat.* **133** (4), 325–336 (1988).
60. Ji, Y., Hao, H., Reynolds, K., McMahon, M. & Zhou, C. J. Wnt signaling in neural crest ontogenesis and oncogenesis. *Cells*, **8**, 10, (2019).
61. Lough, K. J., Byrd, K. M., Spitzer, D. C. & Williams, S. E. Closing the gap: Mouse models to study adhesion in secondary palatogenesis. *J. Dent. Res.* **96** (11), 1210–1220 (2017).
62. Kousa, Y. A. et al. IRF6 and SPRY4 signaling interact in Periderm Development. *J. Dent. Res.* **96** (11), 1306–1313 (2017).
63. G. De La Garza et al., Interferon regulatory factor 6 promotes differentiation of the periderm by activating expression of grainyhead-like 3. *J. Invest. Dermatol.*, **133** (1), 68–77 (2013).
64. Graham, A. Development of the pharyngeal arches. *Am. J. Med. Genet.* **119 A** (3), 251–256 (2003).
65. Liu, N. et al. DNA binding-dependent and -independent functions of the Hand2 transcription factor during mouse embryogenesis. *Development*. **136** (6), 933–942 (2009).
66. Coffin Talbot, J., Johnson, S. L. & Kimmel, C. B. hand2 and dlx genes specify dorsal, intermediate and ventral domains within zebrafish pharyngeal arches. *Development*. **137** (15), 2507–2517 (2010).
67. Thomason, H. A. et al. Cooperation between the transcription factors p63 and IRF6 is essential to prevent cleft palate in mice. *J. Clin. Invest.* **120** (5), 1561–1569 (2010).
68. Lin-Shiao, E. et al. P63 establishes epithelial enhancers at critical craniofacial development genes. *Sci. Adv.* **5** (5), 1–15 (2019).

69. Sun, J. et al. Genetic association and functional validation of ZFP36L2 in non-syndromic orofacial cleft subtypes. *J. Hum. Genet.* **69**, 3–4 (2024).
70. Gianferante, D. M. et al. Genotype-phenotype association and variant characterization in Diamond Blackfan anemia caused by pathogenic variants in RPL35A. *Haematologica*. **106** (5), 1303–1310 (2021).
71. Chiquet, B. T. et al. Variation in WNT genes is associated with non-syndromic cleft lip with or without cleft palate. *Hum. Mol. Genet.* **17** (14), 2212–2218 (2008).
72. Nikopensius, T. et al. Genetic variants in COL2A1, COL11A2, and IRF6 contribute risk to nonsyndromic cleft palate. *Birth Defects Res. Part. - Clin. Mol. Teratol.* **88** (9), 748–756 (2010).
73. Nikopensius, T. et al. Variation in FGF1, FOXE1, and TIMP2genes is associated with nonsyndromic cleft lip with or without cleft palate. *Birth Defects Res. Part. - Clin. Mol. Teratol.* **91** (4), 218–225 (2011).
74. Mostowska, A. et al. Genotype and haplotype analysis of WNT genes in non-syndromic cleft lip with or without cleft palate. *Eur. J. Oral Sci.* **120** (1), 1–8 (2012).
75. Feng, C. et al. C392T polymorphism of the < em > Wnt10a gene in non-syndromic oral cleft in a northeastern Chinese population, Br. J. Oral Maxillofac. Surg., vol. 52, no. 8, pp. 751–755, Oct. (2014).
76. Lu, Y. et al. Variations in WNT3 gene are associated with incidence of non-syndromic cleft lip with or without cleft palate in a northeast Chinese population. *Genet. Mol. Res.* **14** (4), 12646–12653 (2015).
77. Pengelly, R. J. et al. Deleterious coding variants in multi-case families with non-syndromic cleft lip and/or palate phenotypes. *Sci. Rep.* **6** (November 2015), 1–8 (2016).
78. Zieger, H. K. et al. Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of de novo mutations. *Hum. Genet. Genomics Adv.* **4** (1), 100166 (2023).

Author contributions

Conceptualization: A.S., K.U.L.; Analysis and curation of data: A.S., S.H.; Investigation: A.S., S.H., K.U.L.; Resources: E.M., K.U.L; Writing - original draft preparation: A.S., K.U.L.; Writing - review and editing: A.S., S.H., E.M., K.U.L.; All authors read and approved the final manuscript.

Funding

K.U.L is member of the Cluster of Excellence ImmunoSensation - EXC2151–390873048, funded by Deutsche Forschungsgemeinschaft (DFG), and has received support from the DFG (LU 1944/3 – 1).

Open Access funding enabled and organized by Projekt DEAL.

Declarations

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/s41598-024-77724-9.

Correspondence and requests for materials should be addressed to A.S. or K.U.L.

Reprints and permissions information is available at www.nature.com/reprints.

Acknowledgements.
The authors thank Friederike David, Carina Mathey, and Sabrina Henne for technical support during data analysis. The authors gratefully acknowledge the granting of access to the Bonna cluster, which is hosted by the University of Bonn.

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

5 Discussion

In order to elucidate nsCL/P development, many studies have aimed at uncovering genetic risk factors and, subsequently, tried to link genes located at these risk loci to disease pathogenesis. In recent years, this has led to the identification of more than 45 associated regions and considerable progress regarding the elucidation of candidate genes and pathways involved (e.g. Ludwig et al., 2017; Welzenbach et al., 2021). The molecular functions of these candidate genes have been investigated in functional studies using a variety of cell culture, mouse and zebrafish models (Cox et al., 2018; Kousa et al., 2019; Carroll et al., 2020; Lee et al., 2020). While bulk RNA sequencing has often been used to study gene expression, the investigation of developing tissues undergoing dynamic differentiation processes requires gene expression analysis at the single-cell level. Publicly available scRNA-seq data relevant to craniofacial development were scarce until 2019, when two murine scRNA-seq datasets were among the first to allow the analysis of high-resolution expression data from the developing head (Cao et al., 2019; Li et al., 2019). Over the past years, single-cell technologies have been increasingly applied to investigate gene expression patterns and cell types during craniofacial development (Yuan et al., 2020; Ozekin et al., 2023; Yankee et al., 2023; Itai et al., 2024).

Leveraging these recent advancements and available datasets, the present thesis aimed to explore candidate gene expression patterns in scRNA-seq data to identify cell types involved in nsCL/P etiology. For this purpose, the murine scRNA-seq data from Cao *et al.* 2019 and Li *et al*. 2019 were re-analyzed and used to examine the expression patterns of nsCL/P GWAS candidate genes at the level of individual genes (Siewert et al., 2023). Overall, this analysis showed that most candidate genes fall into one of two groups, genes that are predominantly expressed in epithelial cells or genes that are predominantly expressed in mesenchymal-like cell types. Additionally, candidate genes that are known to act together in a genetic pathway, like *IRF6* and *TFAP2A*, belonged to the same of these two groups (Kousa et al., 2019). The analysis further confirmed the gene expression patterns of several candidate genes previously described in functional studies, such as *IRF6*, *TFAP2A*, *ESRP1*, *KRT8*, *KRT18*, *FOXE1*, and *FGFR1* (Bachler and Neubüser, 2001; Moll et al., 2008; Moreno et al., 2009; Kousa et al., 2019; Lee et al., 2020). This

indicates the robustness of scRNA-seq for the examination of gene expression profiles. The re-analyzed murine scRNA-seq data sets were also used to investigate the expression patterns of transcription factors whose binding regions are potentially affected by non-coding *de novo* mutations found in whole-genome sequencing data of individuals with nsCL/P (Zieger et al., 2023). For example, *MSC*, a gene encoding for a transcription factor involved in muscle development in the face (Rosero Salazar et al., 2020), showed specific expression in myocytes, connective tissue progenitors, sensory neurons, palatal epithelium and the maxillary prominences, supporting additional evidence for the involvement of muscle development in nsCL/P etiology. Together, these studies demonstrated the potential of findings from genetic studies to elucidate the underlying biological mechanisms of nsCL/P using scRNA-seq data.

Ongoing advances in single-cell technologies and bioinformatics tools continue to offer novel analysis opportunities. The combination of different data modalities in single-cell multi-omics approaches provides a more comprehensive view of biological systems, for example by identifying more complex cell states and regulatory networks (Yan et al., 2024). For example, single-cell multi-omics sequencing techniques combining data from scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing have identified mesenchymal and ectodermal sub-cell types involved in the fusion process of the upper lip and primary palate, as well as cell type-specific regulatory genes that are crucial for this process in mice (Cai and Yin, 2024). Furthermore, these techniques have also been used to identify key lineage-determining transcription factors for murine secondary palate development (Yan et al., 2024). In addition to multi-omics sequencing techniques, bioinformatics approaches offer the possibility of multimodal analysis by combining different data types in one statistical analysis. This potentially facilitates functional interpretation of results obtained from GWAS and other genetic studies, which is crucial for ultimately identifying causal variants and understanding of the underlying biology and pathobiology. For instance, methods such as the single-cell disease relevance score (scDRS), which combines GWAS results and scRNA-seq data, enable polygenic gene expression analysis at single-cell resolution (Zhang et al., 2022). This allows the identification of associations between groups of GWAS candidate genes and single cells.

With the aim to further deepen our understanding of nsCL/P and to build on the previous work, these technological advances have been utilized in the present thesis through the implementation of scDRS (Siewert et al., 2024). The earlier work on murine scRNA-seq data was complemented by an analysis using scRNA-seq data from the heads of human embryos, which only became available in 2023, and nsCL/P GWAS results (Welzenbach et al., 2021; Xu et al., 2023). In contrast to the previous work investigating expression patterns gene by gene, this analysis jointly examined the expression of a group of candidate genes, reflecting the polygenic nature of nsCL/P. This scDRS analysis found epithelial cells and *HAND2*+ pharyngeal arches, a specific pharyngeal arch subcluster, to be associated with the nsCL/P GWAS candidate genes (Siewert et al., 2024). Additionally, the epithelial cells showed heterogeneity regarding this association, indicating a distinct epithelial sub cell type that is involved in nsCL/P. Compared to our previous studies, which examined gene expression in an exploratory manner, the scDRS analysis was able to identify nsCL/P-associated cell types on a more systematic level by combining gene expression levels of all candidate genes into one joint association score. However, the identified associated cell types were consistent with the two groups of expression patterns from the previous study in mice (Siewert et al., 2023). Subsequently, co-expression network analyses in these cell types revealed genes that are known interaction partners, like *IRF6* and *TFAP2A* as well as *IRF6* and *TP63*, to be in the same network (Thomason et al., 2010; Kousa et al., 2019). By combining these co-expression networks with summary statistics from the initial nsCL/P GWAS data from Welzenbach *et al*. 2021, two novel candidate genes were identified, *BFAR* and *HYAL2*. Moreover, supporting evidence for two previously described candidate genes, *CTNND1* and *PRTG*, was found (Leslie et al., 2017; Cox et al., 2018). Overall, this approach represents a novel strategy for candidate gene identification that combines genetic and single-cell transcriptomic data. Taken together, this study demonstrates the potential that lies in the combination of different modalities such as scRNA-seq data and findings from genetic studies to elucidate the underlying biological mechanisms of developmental diseases such as nsCL/P.

However, despite these advantages and constant developments, scRNA-seq data still have considerable technical limitations. For instance, an incomplete detection of molecules in some cells might cause misleading results that do not accurately reflect the true biological meaning (Sreenivasan et al., 2022). Additionally, differences in the handling

of samples prior to sequencing may lead to batch effects between samples. In scRNA-seq data analysis, these can strongly influence the outcome of downstream analysis steps like dimensionality reduction and clustering (Hicks et al., 2018). In Siewert *et al*. 2024, batch effects were minimized between the samples in the human scRNA-seq data set by integrating the samples based on shared cell populations across samples (Stuart et al., 2019). Based on the results obtained in this thesis, future studies on nsCL/P development may specifically concentrate on the identified nsCL/P-associated cell types in functional experiments. More precisely, spatially resolved transcriptomics will allow the investigation of gene expression patterns, while preserving the spatial integrity of developing tissues and underlying cell-cell signaling processes (Tseng and Crump, 2023). Additionally, methods such as RNA velocity and lineage tracing may uncover dynamic processes during craniofacial development that potentially result in nsCL/P (Wagner and Klein, 2020; Wang et al., 2024).

In summary, by combining genetic and single-cell transcriptomic data, this thesis identified developmental cell types with a potential role in nsCL/P. Moreover, the systematic approach led to the identification of novel nsCL/P candidate genes, while the results in general constitute a substantiated basis for future functional analyses.

REFERENCES

Bachler, M., and Neubüser, A. (2001). Expression of members of the Fgf family and their receptors during midfacial development. *Mech Dev* 100, 313–316

Cai, S., and Yin, N. (2024). Single-cell transcriptome and chromatin accessibility mapping of upper lip and primary palate fusion. *J Cell Mol Med* 28, e70128

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 496–502

Carroll, S. H., Trevino, C. M., Li, E. B., Kawasaki, K., Myers, N., Hallett, S. A., et al. (2020). An Irf6-Esrp1/2 regulatory axis controls midface morphogenesis in vertebrates. *Development (Cambridge)* 147

Cox, L. L., Cox, T. C., Moreno Uribe, L. M., Zhu, Y., Richter, C. T., Nidey, N., et al. (2018). Mutations in the Epithelial Cadherin-p120-Catenin Complex Cause Mendelian Non-Syndromic Cleft Lip with or without Cleft Palate. *Am J Hum Genet* 102, 1143–1157

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578

Itai, T., Yan, F., Liu, A., Dai, Y., Iwaya, C., Curtis, S. W., et al. (2024). Investigating gene functions and single-cell expression profiles of de novo variants in orofacial clefts. *Human Genetics and Genomics Advances* 5, 100313

Kousa, Y. A., Zhu, H., Fakhouri, W. D., Lei, Y., Kinoshita, A., Roushangar, R. R., et al. (2019). The TFAP2A-IRF6-GRHL3 genetic pathway is conserved in neurulation. *Hum Mol Genet* 28, 1726–1737

Lee, S. K., Sears, M. J., Zhang, Z., Li, H., Salhab, I., Krebs, P., et al. (2020). Cleft lip and cleft palate in Esrp1 knockout mice is associated with alterations in epithelial-mesenchymal crosstalk. *Development (Cambridge)* 147

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Buxó, C. J., Castilla, E. E., Christensen, K., et al. (2017). Association studies of low-frequency coding variants in nonsyndromic cleft lip with or without cleft palate. *Am J Med Genet A* 173, 1531–1538

Li, H., Jones, K. L., Hooper, J. E., and Williams, T. (2019). The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. *Development (Cambridge)* 146

Ludwig, K. U., Böhmer, A. C., Bowes, J., Nikolić, M., Ishorst, N., Wyatt, N., et al. (2017). Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. *Hum Mol Genet* 26, 829–842

Moll, R., Divo, M., and Langbein, L. (2008). The human keratins: Biology and pathology. *Histochem Cell Biol* 129, 705–733

Moreno, L. M., Mansilla, M. A., Bullard, S. A., Cooper, M. E., Busch, T. D., Machida, J., et al. (2009). FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Hum Mol Genet* 18, 4879–4896

Ozekin, Y. H., O'Rourke, R., and Bates, E. A. (2023). Single cell sequencing of the mouse anterior palate reveals mesenchymal heterogeneity. *Developmental Dynamics* 252, 713–727

Rosero Salazar, D. H., Carvajal Monroy, P. L., Wagener, F. A. D. T. G., and Von den Hoff, J. W. (2020). Orofacial Muscles: Embryonic Development and Regeneration after Injury. *J Dent Res* 99, 125–132

Siewert, A., Hoeland, S., Mangold, E., and Ludwig, K. U. (2024). Combining genetic and single-cell expression data reveals cell types and novel candidate genes for orofacial clefting. *Sci Rep* 14, 26492

Siewert, A., Reiz, B., Krug, C., Heggemann, J., Mangold, E., Dickten, H., et al. (2023). Analysis of candidate genes for cleft lip ± cleft palate using murine single-cell expression data. *Front Cell Dev Biol* 11, 1–11

Sreenivasan, V. K. A., Balachandran, S., and Spielmann, M. (2022). The role of single-cell genomics in human genetics. *J Med Genet* 59, 827–839

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21

Thomason, H. A., Zhou, H., Kouwenhoven, E. N., Dotto, G. P., Restivo, G., Nguyen, B. C., et al. (2010). Cooperation between the transcription factors p63 and IRF6 is essential to prevent cleft palate in mice. *Journal of Clinical Investigation* 120, 1561–1569

Tseng, K.-C., and Crump, J. G. (2023). Craniofacial developmental biology in the single-cell era. *Development* 150

Wagner, D. E., and Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nat Rev Genet* 21, 410–427

Wang, B., Zhang, Z., Zhao, J., Ma, Y., Wang, Y., Yin, N., et al. (2024). Spatiotemporal Evolution of Developing Palate in Mice. *J Dent Res* 103, 546–554

Welzenbach, J., Hammond, N. L., Nikolić, M., Thieme, F., Ishorst, N., Leslie, E. J., et al. (2021). Integrative approaches generate insights into the architecture of non-syndromic cleft lip with or without cleft palate. *Human Genetics and Genomics Advances* 2, 1–14

Xu, Y., Zhang, T., Zhou, Q., Hu, M., Qi, Y., Xue, Y., et al. (2023). A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nat Cell Biol*, 1–12

Yan, F., Suzuki, A., Iwaya, C., Pei, G., Chen, X., Yoshioka, H., et al. (2024). Single-cell multiomics decodes regulatory programs for mouse secondary palate development. *Nat Commun* 15, 1–17

Yankee, T. N., Oh, S., Winchester, E. W., Wilderman, A., Robinson, K., Gordon, T., et al. (2023). Integrative analysis of transcriptome dynamics during human craniofacial development identifies candidate disease genes. *Nat Commun* 14, 4623

Yuan, Y., Loh, Y. H. E., Han, X., Feng, J., Ho, T. V., He, J., et al. (2020). Spatiotemporal cellular movement and fate decisions during first pharyngeal arch morphogenesis. *Sci Adv* 6

Zhang, M. J., Hou, K., Dey, K. K., Sakaue, S., Jagadeesh, K. A., Weinand, K., et al. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat Genet*

Zieger, H. K., Weinhold, L., Schmidt, A., Holtgrewe, M., Juranek, S. A., Siewert, A., et al. (2023). Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of de novo mutations. *Human Genetics and Genomics Advances* 4, 100166

6 Acknowledgment

First and foremost, I would like to thank Prof. Dr. Kerstin Ludwig for giving me the opportunity to pursue my PhD in her group and providing me with the interesting research topic. I sincerely thank you for your kindness, your guidance and the many laughs we have shared throughout the years.

I would also like to thank Prof. Dr. Jan Hasenauer for taking over as my second reviewer, Prof. Dr. Hubert Schorle and Prof. Dr. Oliver Gruß as members of my thesis committee, as well as Prof. Dr. Peter Krawitz as a former member of my thesis committee, for their time and consideration.

I would like to thank Dr. Elisabeth Mangold and Dr. Julia Heggemann for their cooperation and feedback on various projects over the past years. I want to thank all former and current members of the Ludwig lab, especially Ronja, Hanna, Axel, Madhu and Ayda, for their support and the fun times we had at and outside of work.

Further, I would like to thank Dr. Benedikt Reiz and Dr. Henning Dickten for a rewarding collaboration.

I want to thank all colleagues at the Life & Brain and the Institute of Human Genetics for their support, the friendly cooperation and especially the fun lunch breaks, forest walks and Karneval celebrations. A very special thank you to Carina, Rike, Charlotte, Anna So and Julia for your friendship, moral support, impressive baking skills, our gym sessions and the countless funny moments I got to share with you over the last few years. I will treasure this time forever!

I would like to thank all my friends who have accompanied and encouraged me through the ups and downs of the last few years. A special thanks to my friend Isabell for always cheering me on.

My deepest gratitude goes to my partner Tobi and my family, especially Mama, Flo, my grandparents and my favorite person in the universe, my twin sister Alina. Thank you for your love, support and constant encouragement.