

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 *ZFX4* (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

¹Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn 53127, Germany; ²Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn 53127, Germany; ³Core Unit Bioinformatics, Berlin Institute of Health, Berlin 10117, Germany; ⁴Department of Oncology, Hematology and Rheumatology, University Hospital Bonn, Bonn 53127, Germany; ⁵Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn 53127, Germany; ⁶Max Delbrück Center for Molecular Medicine, Berlin 13125, Germany

*Correspondence: kerstin.ludwig@uni-bonn.de
<https://doi.org/10.1016/j.xhgg.2022.100166>

© 2022 The Authors. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



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 GMKF. 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 ncER²¹ ([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 Predictor²² (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 browser²⁸ ([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 described⁴ (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.

^aExonic DNMs based on GENCODE Basic gene annotation version33.hg19, including non-coding parts of gene sequences (e.g., 3'/5' UTRs).

^bEffect combinations from Variant Effect Predictor output were reduced to classes (see Table S4 for grouped effect names). LoF DNMs include nonsense and frameshift DNMs.

^cIn 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 p³² 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 ≥ 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 ≥ 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 families⁶; thereby, this result further supports the role of the PLEKHA-family in nsCL/P etiology.

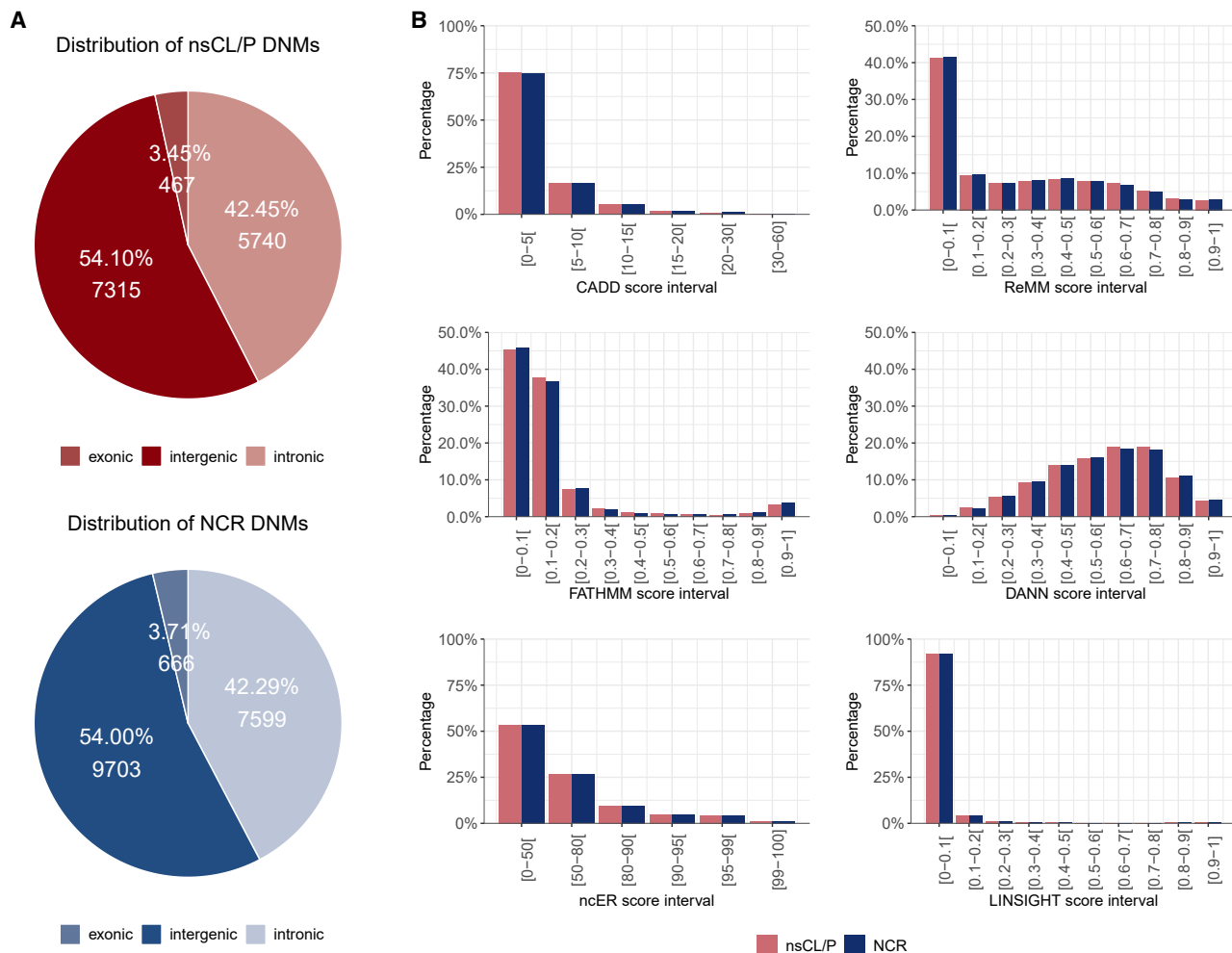


Figure 1. Comparative analyses of *de novo* mutations

(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 tissue²⁶ (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

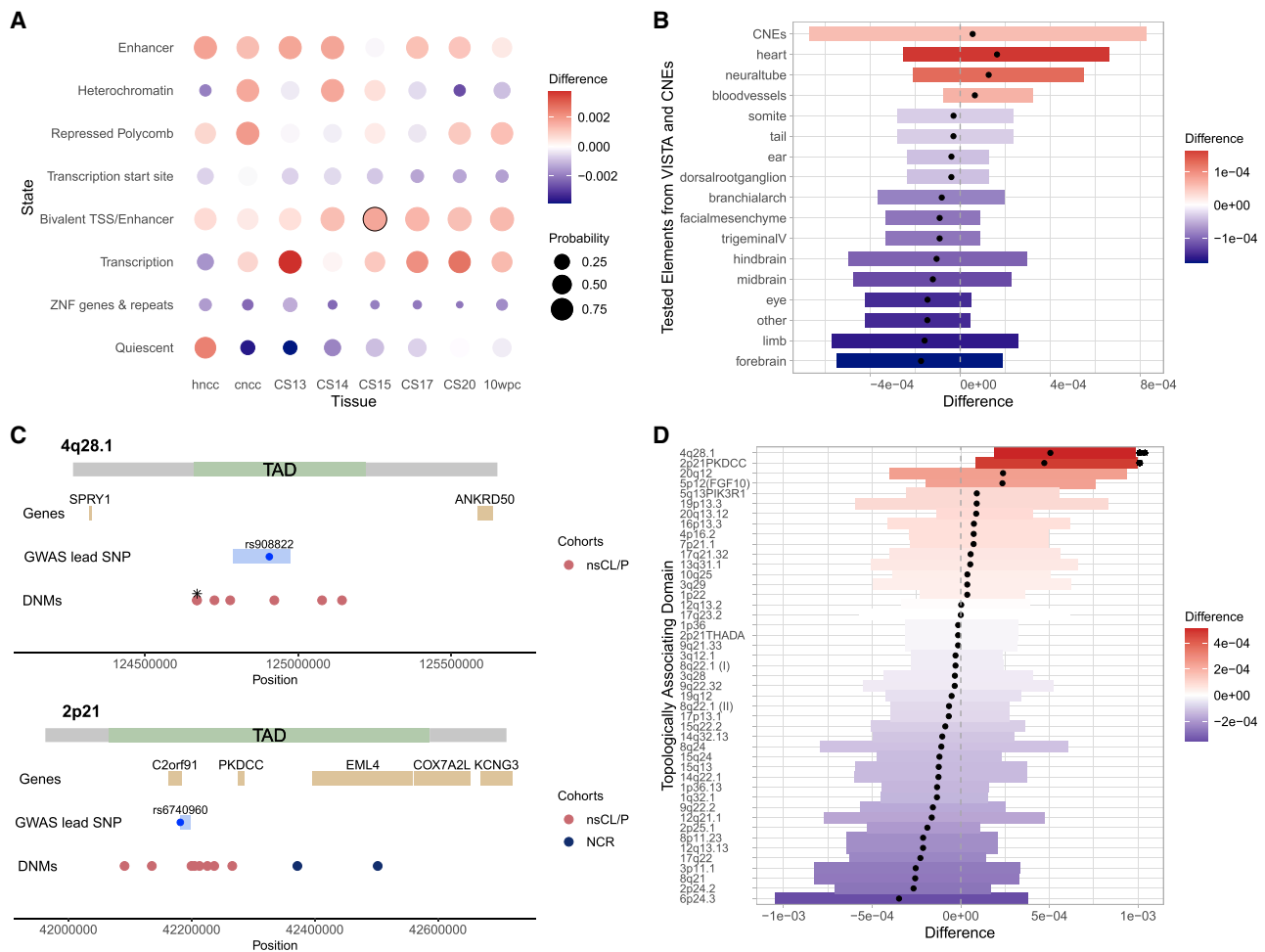


Figure 2. Enrichment of non-syndromic cleft lip with/without cleft palate *de novo* mutations in genomic candidate regions

(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 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 4q28.1⁴⁰ (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

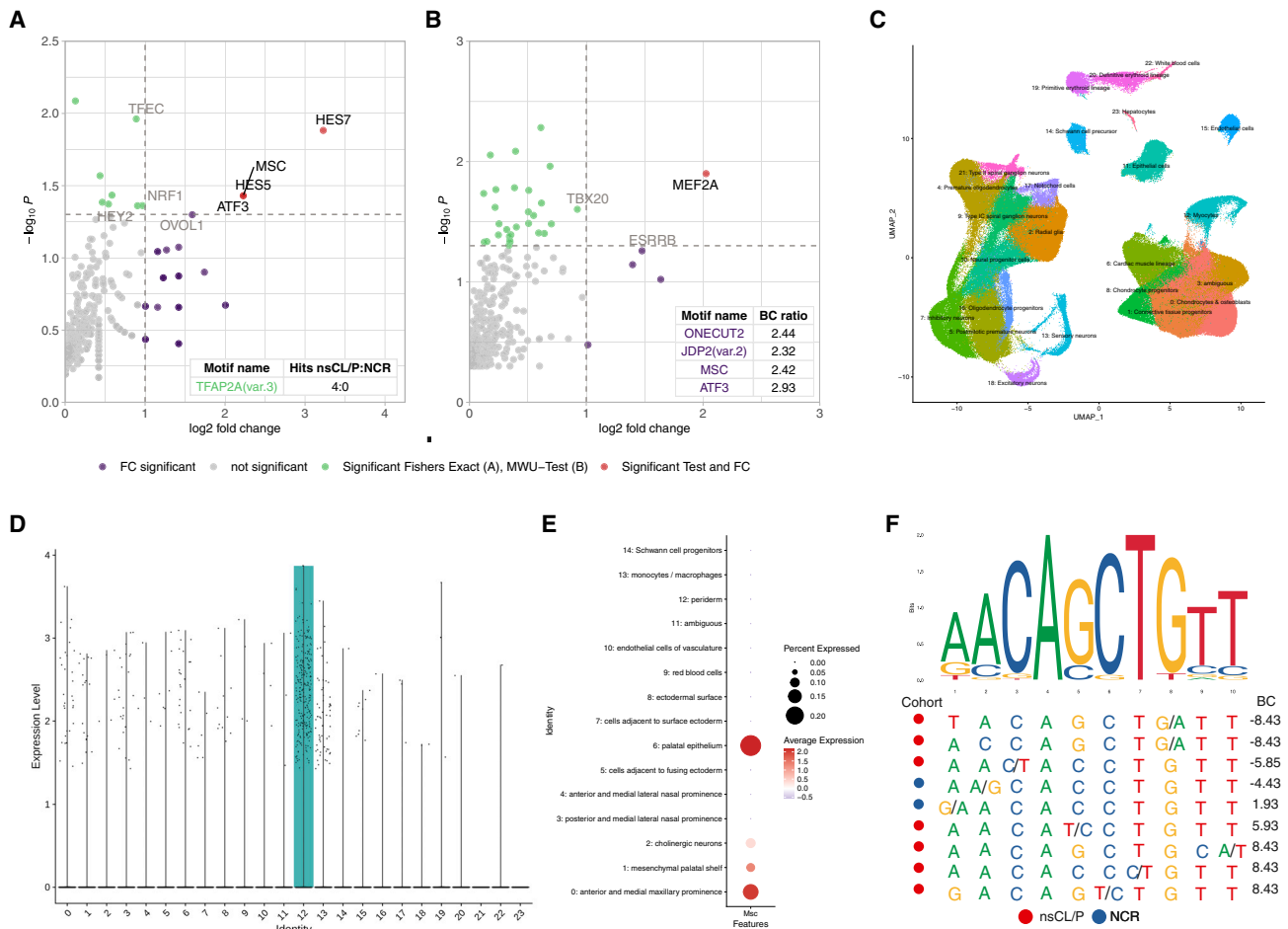


Figure 3. Identification of Musculin as a player in non-syndromic cleft lip with/without cleft palate etiology

(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 \log_2 FC (on y axis) versus Fisher's exact tests ($-\log_{10}(p \text{ value})$ on x axis). Insert represents motif TFAP2a (var.3) that had \log_2 FC ≥ 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 \log_2 FC (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 \log_2 FC ≥ 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/MS; all $p < 0.05$), and a \log_2 FC ≥ 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 $\log_2\text{FC} \geq 1$, 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 (MOCA⁴³; 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, *ZFX4*,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 lambdaoid 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 Donnell 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.

Web resources

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

denovoLOBOG, <https://github.com/pjshort/denovoTF>.
FunciVar, <https://github.com/Simon-Coetzee/funcivar>.
GEO, <https://www.ncbi.nlm.nih.gov/geo/>
GENCODE, https://www.encodegenes.org/human/grc/h37_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

- Mangold, E., Ludwig, K.U., and Nöthen, M.M. (2011). Breakthroughs in the genetics of orofacial clefting. *Trends Mol. Med.* 17, 725–733.
- 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.
- 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.
- 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.
- Basha, M., Demeer, B., Revenu, 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.
- 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.
- 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., GOS-gene, 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.
- 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.
- 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.
- 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.
- Bishop, M.R., Diaz Perez, K.K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J.B., Taub, M.A., Moreno-Urbe, 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.
- 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.
- Cvijetkovic, 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763.
- 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.
- 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., Götz, 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 non-syndromic 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, E., 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.

Supplemental information

**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

Table of Contents

Content	Page(s)
Figure S1: Allele frequencies of all DNMs	Page 2
Figure S2-S4: Quality control	Pages 3-5
Figure S5-S6: nsCL/P phenotype and sex distribution	Page 6
Figure S7: Number of DNMs per trio for nsCL/P and NCR	Page 7
Figures S8-S10: Distribution of prediction scores restricted to non-coding DNMs (8), raw CADD score values for all DNMs (9), and number of DNMs above threshold of multiple scores (10)	Pages 8-10
Figure S11: Number of predicted transcription factor binding sites per DNM.	Page 11
Figure S12-S14: Single-cell data analysis from the Mouse Organogenesis Cell Atlas	Pages 12-14
Figure S15: Single-cell data analysis from the lambdoidal junction for <i>Atf3</i>	Page 15
Figure S16: EMSA experiments	Page 16
Figure S17: Gene network involved in development of orofacial branchiomeric muscles	Page 17
Tables S1-S3: TADs _{GWAS} (1), Genomic sequences tested with EMSA (2), recurrent DNMs (3)	Excel Spreadsheet
Table S4: Grouped variant effects by Variant Effect Predictor	Page 18
Tables S5-S6: Coding nsCL/P DNMs (5), Comparison with DNMs in Bishop et al. (6)	Excel Spreadsheet
Tables S7-S15: Distribution of <i>in silico</i> prediction scores	Tables S7-S14: Pages 19-21
	Table S15: Excel Spreadsheet
Tables S16-26: Element-wise DNM enrichment analyses	Excel Spreadsheet
Tables S27-S30: Analysis of transcription factor binding sites	Tables S27-S28: Excel Spreadsheet
	Tables S29-30: Page 22
Table S31: DNMs in <i>ZFHX4</i>	Page 23
Supplemental Methods	Pages 24-30
References of Supplement	Pages 31-32

Abbreviations: DNMs – de novo mutations; nsCL/P – non-syndromic cleft lip with or without cleft palate; NCR – non-cleft reference; EMSA – electrophoretic mobility shift assay

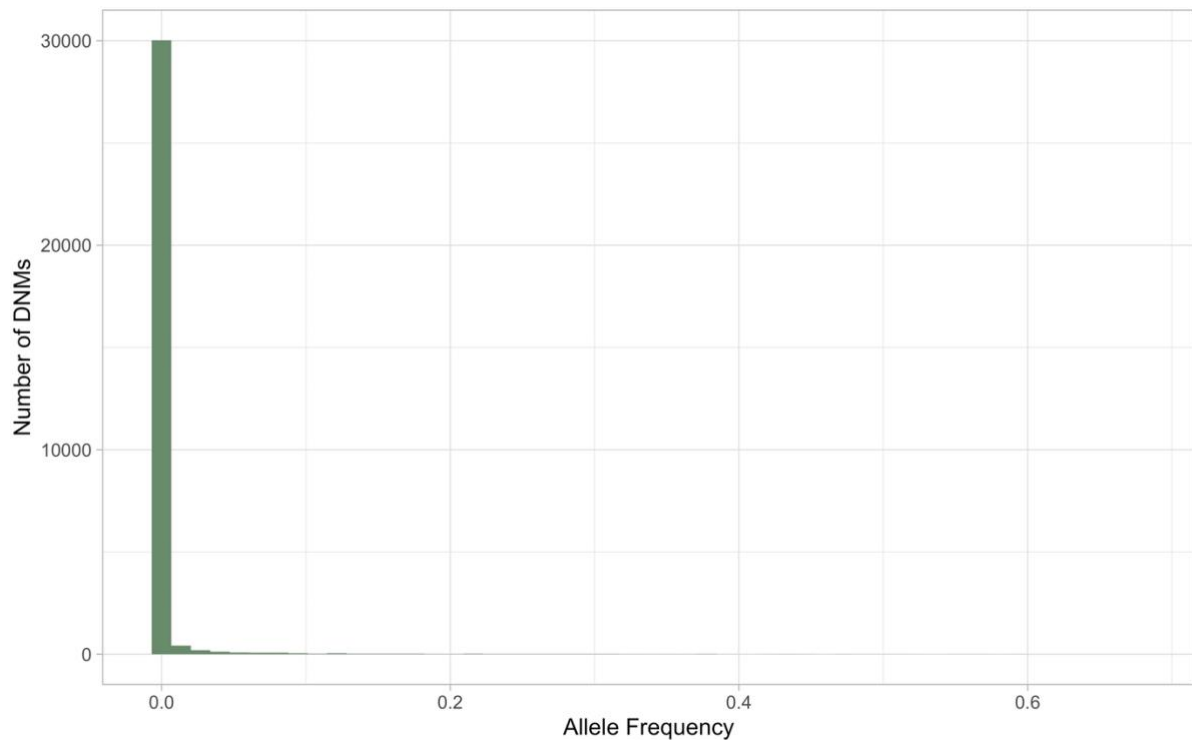


Figure S1. Distribution of allele frequency of all *de novo* mutations in dataset.

Allele frequency for all populations was annotated using gnomAD v3.1.1. The histogram shows the allele frequency of all 31,490 *de novo* mutations (DNMs) from nsCL/P and NCR individuals (binwidth: 0.0135).

Abbreviations: nsCL/P – non-syndromic cleft lip with/without cleft palate; NCR – non-cleft reference

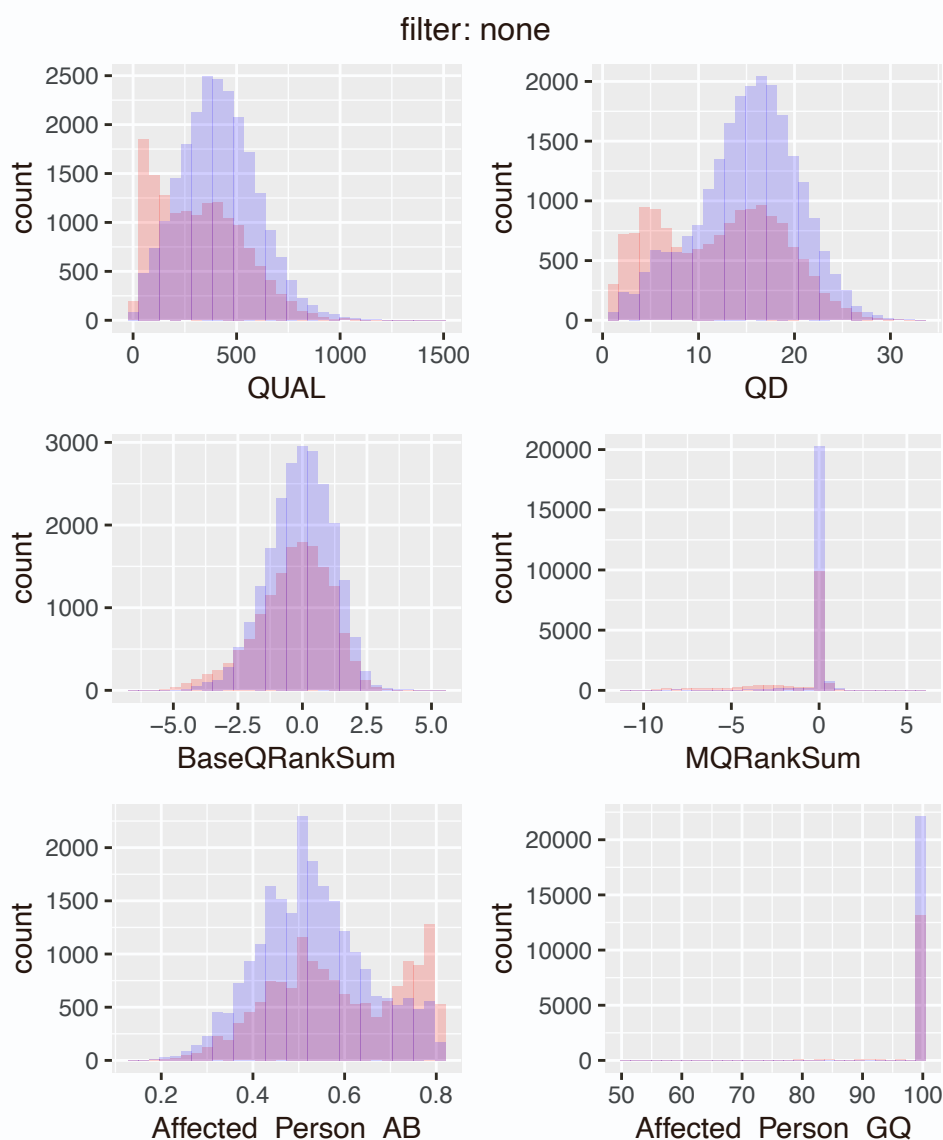


Figure S2. Different quality parameters of *de novo* mutations. The individual histograms show quality scores of *de novo* mutations (post sample QC; intersect between Haplotype Caller and Unified Genotyper). The QUAL, QD, BaseQRankSum, and Affected_Person_GQ values are the values determined for the variant position or variant call for the index patient by the Haplotype Caller. The Affected_Person_AB value corresponds to the allelic balance of the index patient (read count of the alternative allele relative to the total read count). For each histogram, data for known variants (red, variant in gnomAD genomes version 2.0.1, in the 1000 genomes project, or in the Exome Sequencing Project) and non-known variants (blue) are shown overlaid. Note the relative enrichment of known variants in the segments of the histograms with low-quality scores.

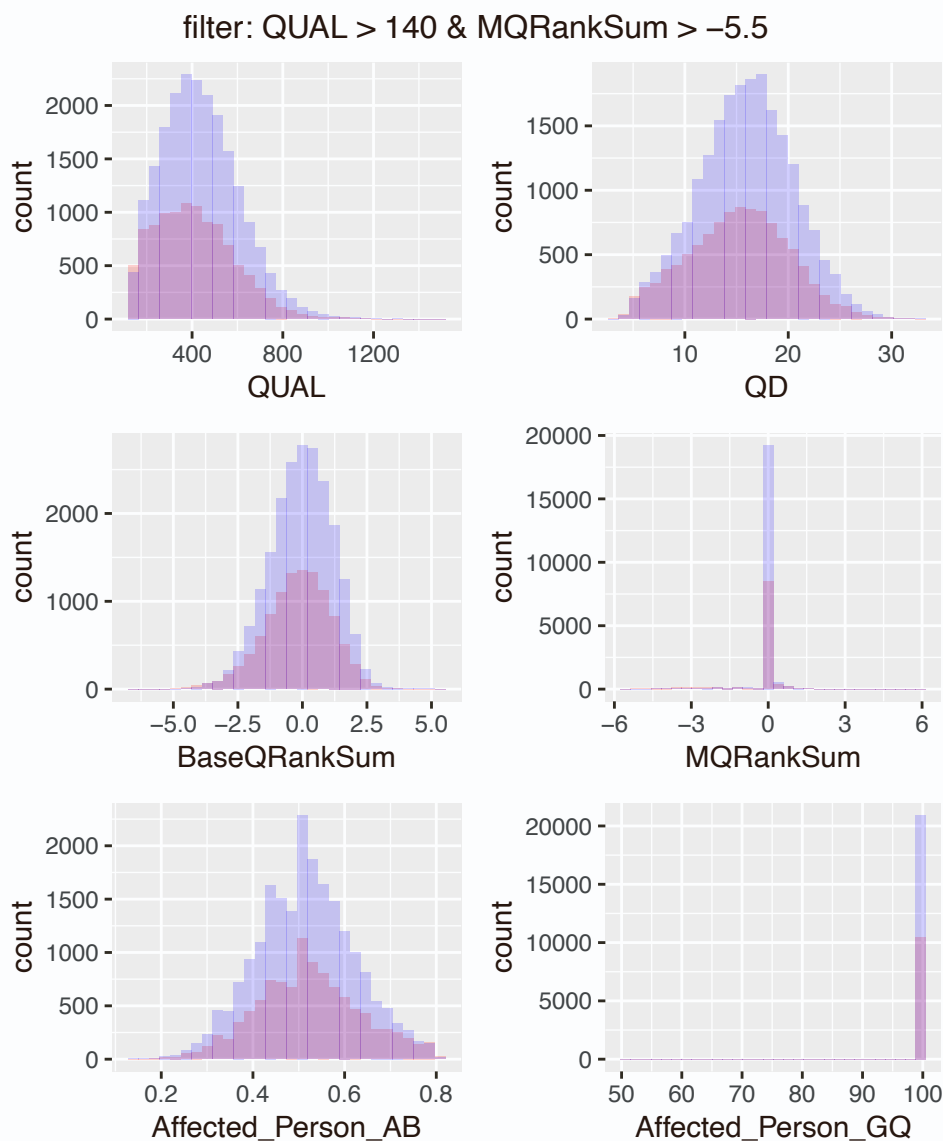


Figure S3. Histograms of quality scores of *de novo* mutations after filtering on QUAL and MQRankSum. Representation analogous to Figure S2: The individual histograms show quality scores of *variants de novo* mutations (post sample QC; intersect between Haplotype Caller and Unified Genotyper). However, *de novo* mutations were filtered for QUAL > 140 and MQRankSum > -5.5. Cut-offs were determined visually using the histograms shown in Figure S2.

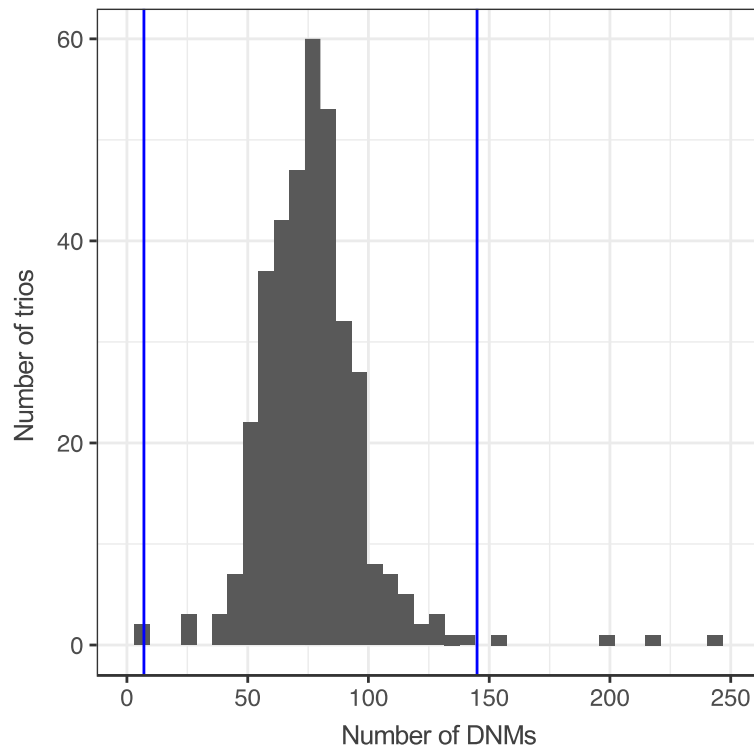


Figure S4. Number of *de novo* mutations per trio. The histogram shows the number of trios with the respective number of *de novo* mutations (DNMs; binwidth: 6.25). Trios with a number of DNMs above median + 3x IQR or below median - 3xIQR (blue line) were excluded for the following analyses. The cut-off was determined visually using the histogram shown. Note that the histogram only shows the range between 0 and 250 DNMs per trio. Therefore, extreme outliers are not shown.

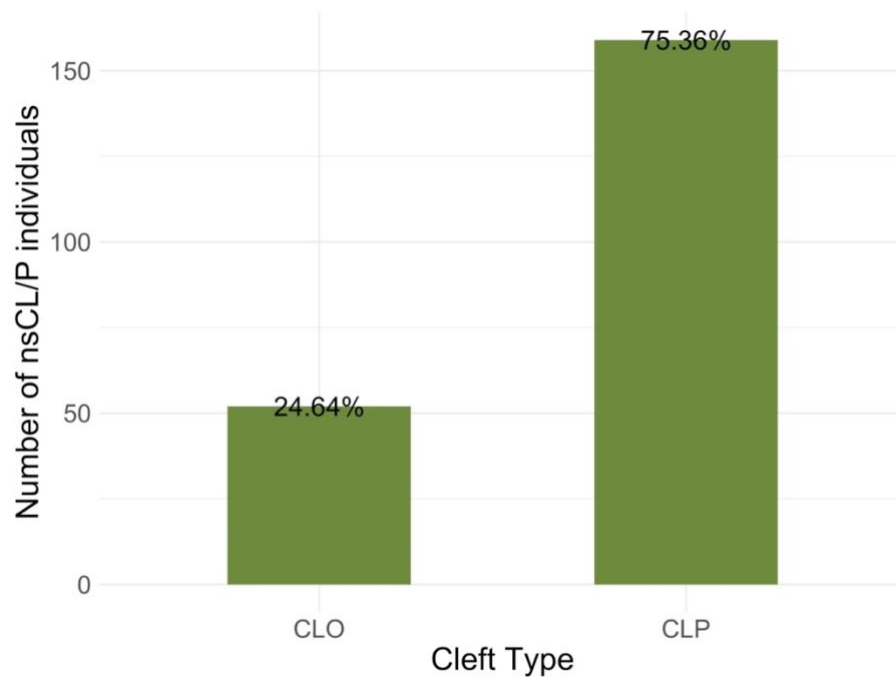


Figure S5. Distribution of different cleft phenotypes in 211 nsCL/P individuals. 52 individuals (24.6%) showed a cleft lip only (CLO) and 159 individuals (75.4%) cleft lip and cleft palate (CLP).
Abbreviation: nsCL/P – non-syndromic cleft lip with or without cleft palate

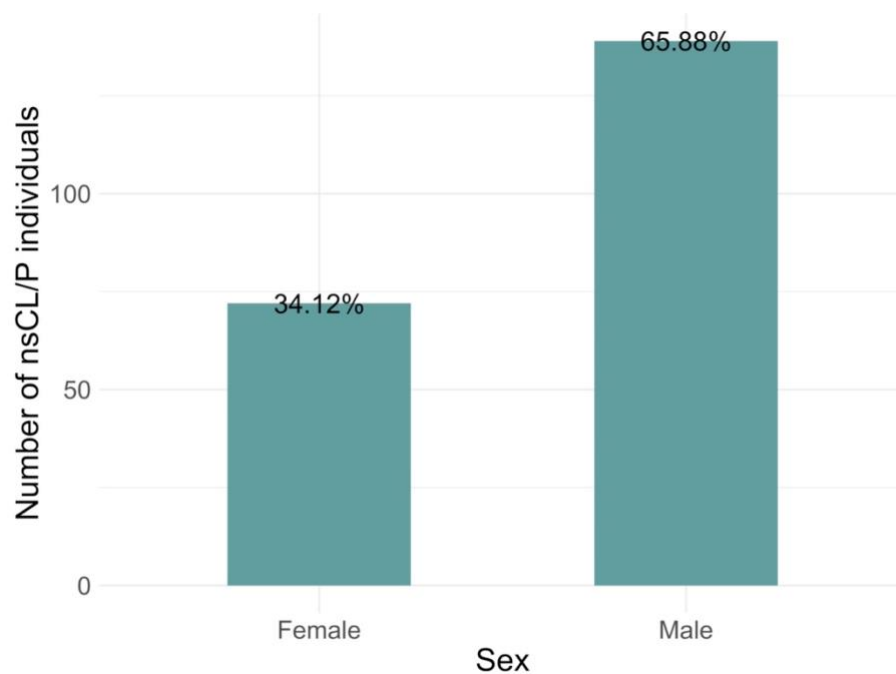


Figure S6. Distribution of sex in 211 nsCL/P individuals.
Abbreviation: nsCL/P – non-syndromic cleft lip with or without cleft palate

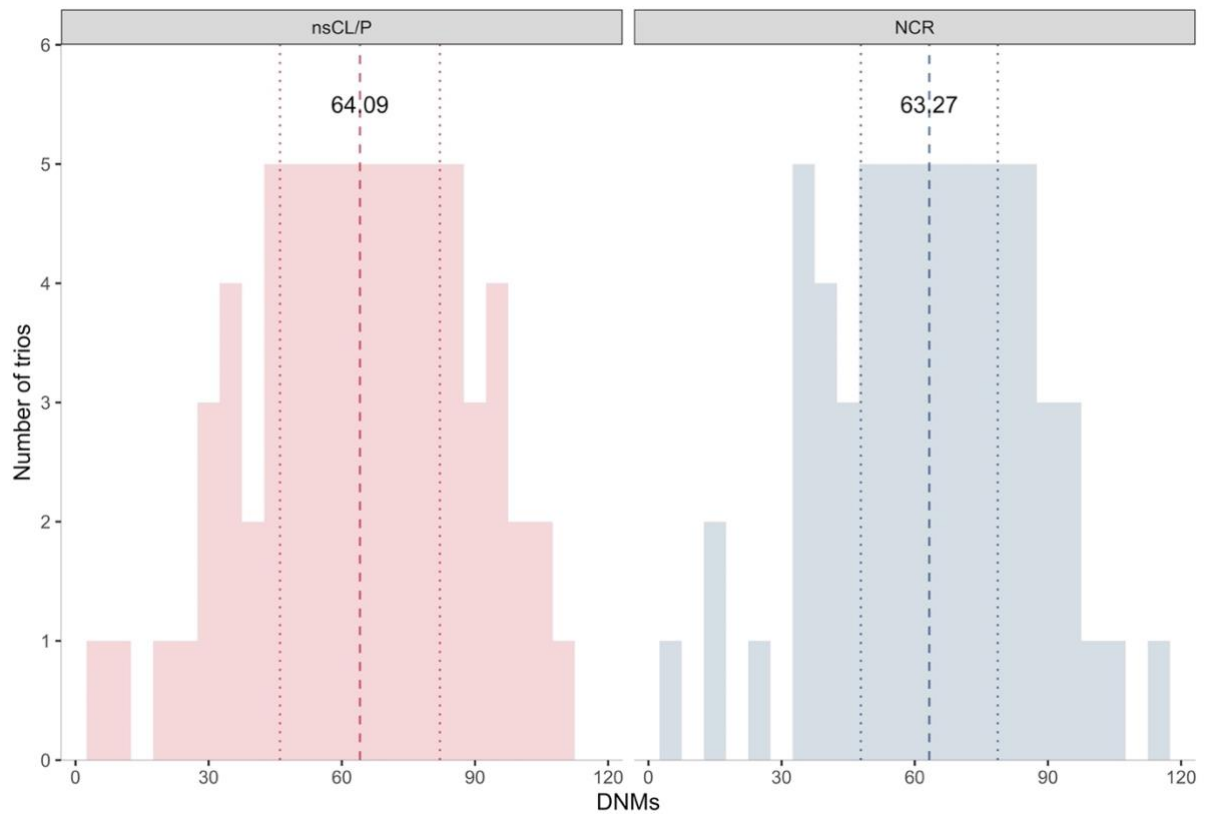


Figure S7. Number of *de novo* mutations per trio. Mean and standard deviation for number of *de novo* mutations (DNMs) are shown by lines (dashed: mean of DNMs per sample in cohorts, dotted: standard deviation of DNMs per sample in cohorts). Binwidth = 5, mean number of DNMs shown over dashed line. Abbreviations: nsCL/P – non-syndromic cleft lip with or without cleft palate; NCR – non-cleft reference cohort

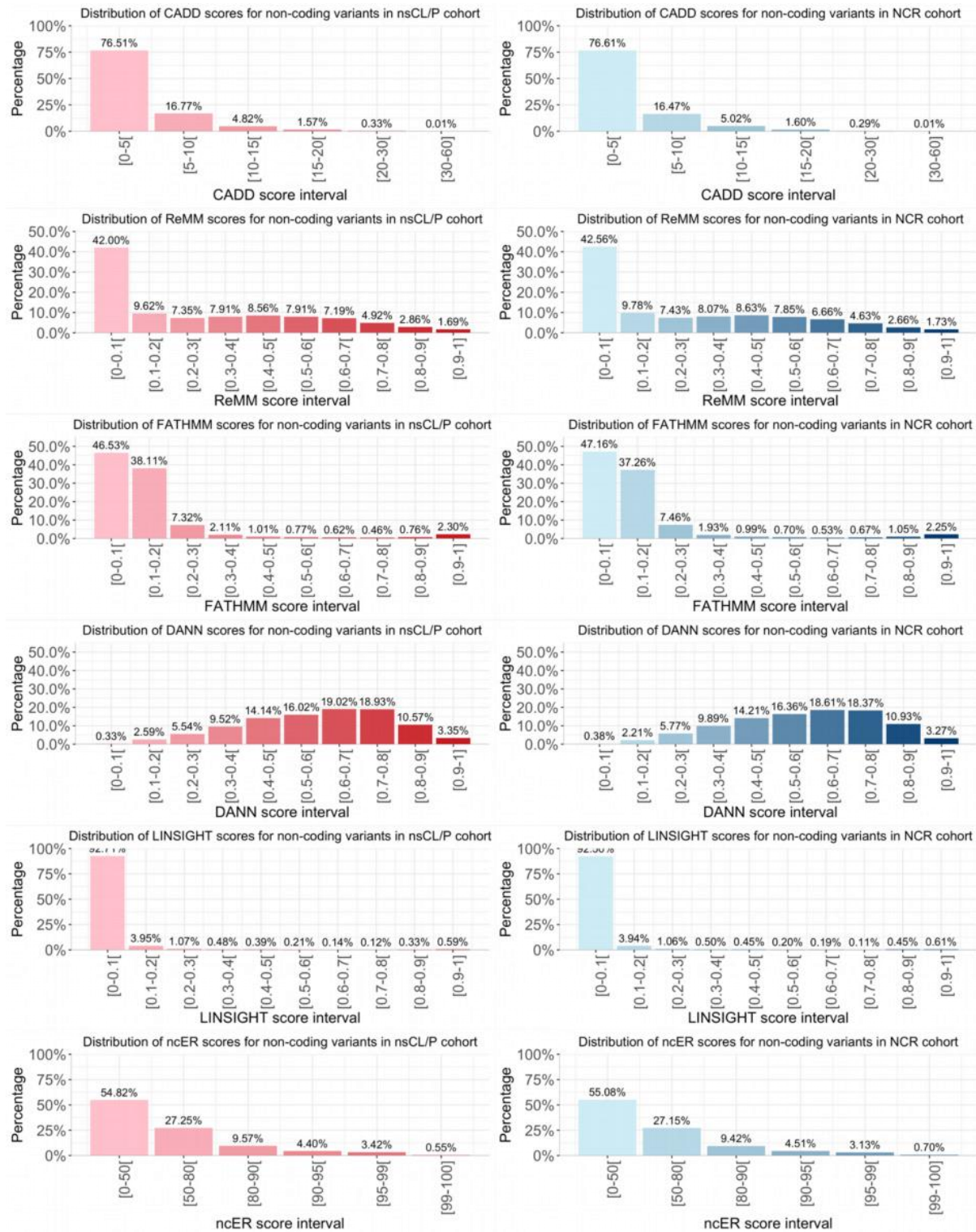


Figure S8. Distribution of prediction scores for non-coding *de novo* mutations in both cohorts. Distribution of six *in silico* prediction scores for non-coding *de novo* mutations (DNMs) in non-syndromic cleft lip with/without cleft palate (nsCL/P; red) and non-cleft reference cohort (NCR; blue). Thresholds and references for six *in silico* prediction scores included are shown in Table S7.

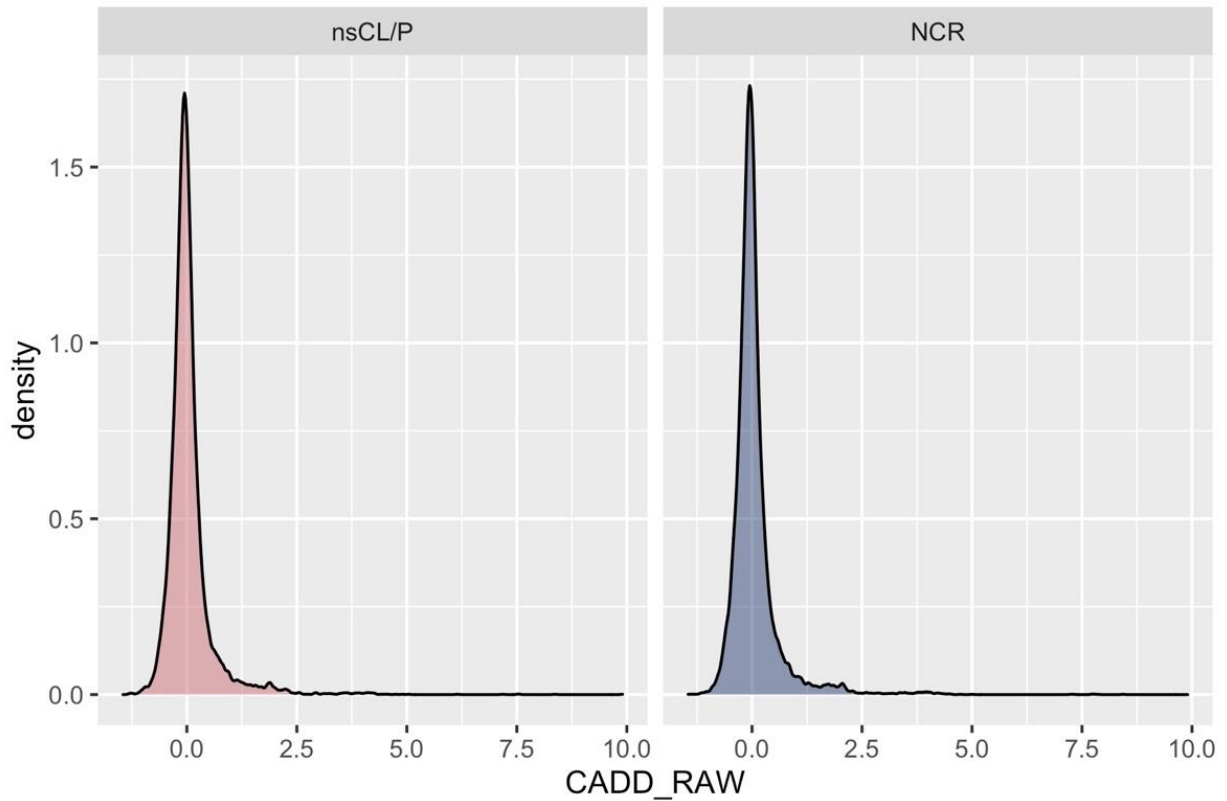


Figure S9. Cohort-wise distribution of raw CADD values for *de novo* mutations.

This density plot shows the distribution of raw CADD values for nsCL/P *de novo* mutations (DNMs) in red and the distribution of raw CADD values for NCR DNMs in blue.

Abbreviations: CADD - Combined Annotation–Dependent Depletion (v1.4, Kircher et al., 2014); nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

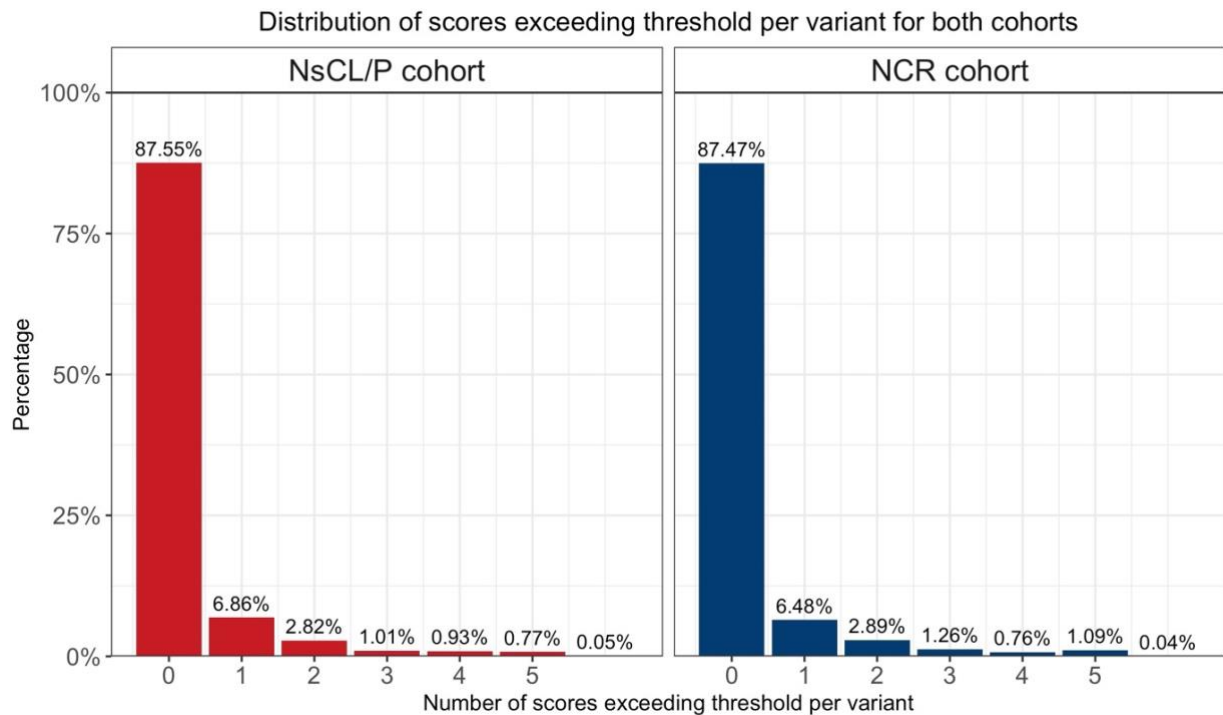


Figure S10. Percentage of *de novo* mutations exceeding the respective thresholds for the indicated number of *in silico* scores. Thresholds and references for six *in silico* prediction scores used for the comparison of cohorts are shown in Table S7.

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

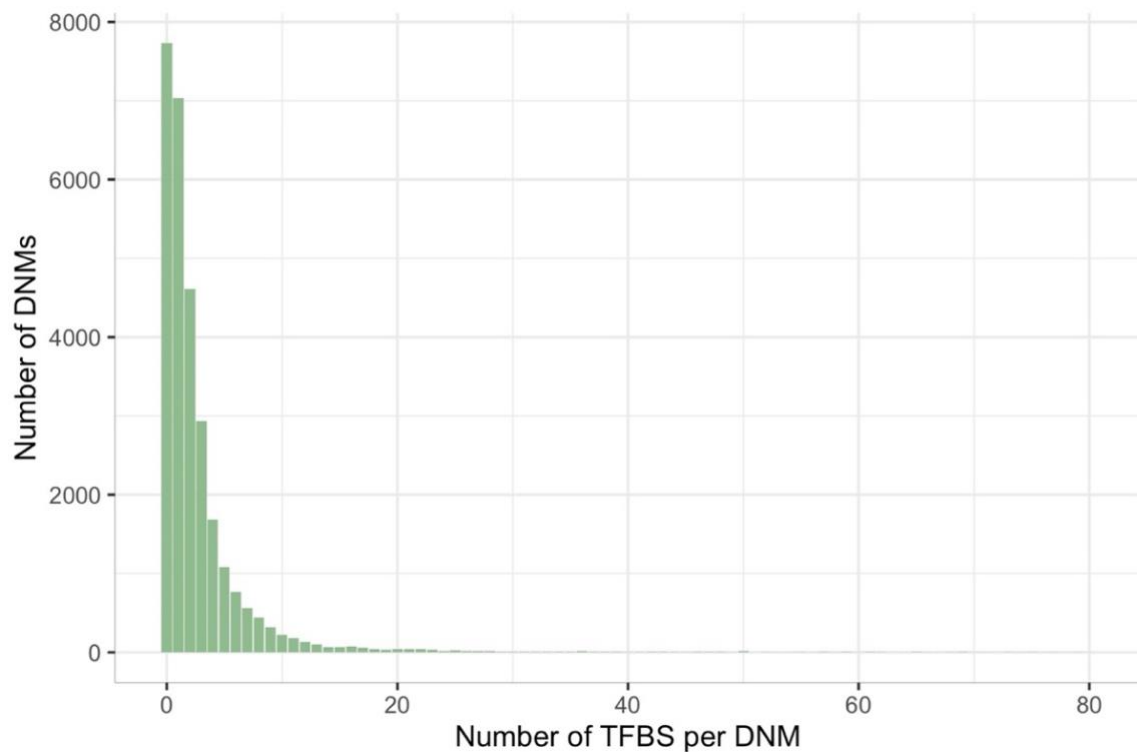


Figure S11. Number of predicted transcription factor binding sites per *de novo* mutation. For transcription factor binding sites (TFBS) identification, position weight matrix (PWM) information was compared to the genomic sequence around each DNM, with reference and alternative allele, using 810 PWMs from Jaspar 2020. For 21,043 out of 28,773 tested DNMs (only single nucleotide substitutions included) transcription factor binding events were detected.
Abbreviations: TFBS – transcription factor binding sites; DNMs - *de novo* mutations

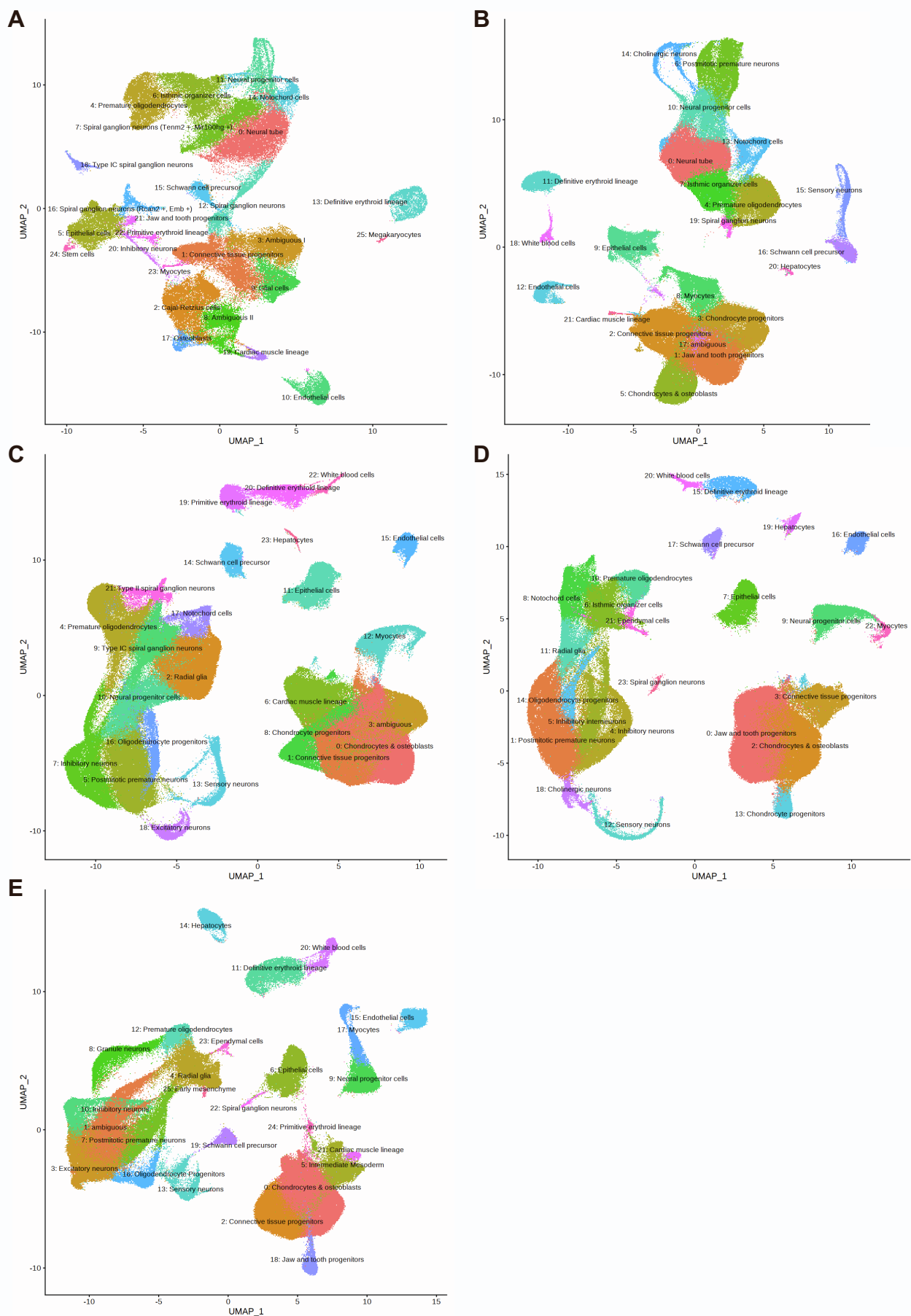


Figure S12. Single-cell data during murine embryogenesis. UMAP plots with cell clusters from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019) for embryonic days (A) E9.5, (B) E10.5, (C) E11.5, (D) E12.5, and (E) E13.5. The annotation of cell clusters is based on the original publication.

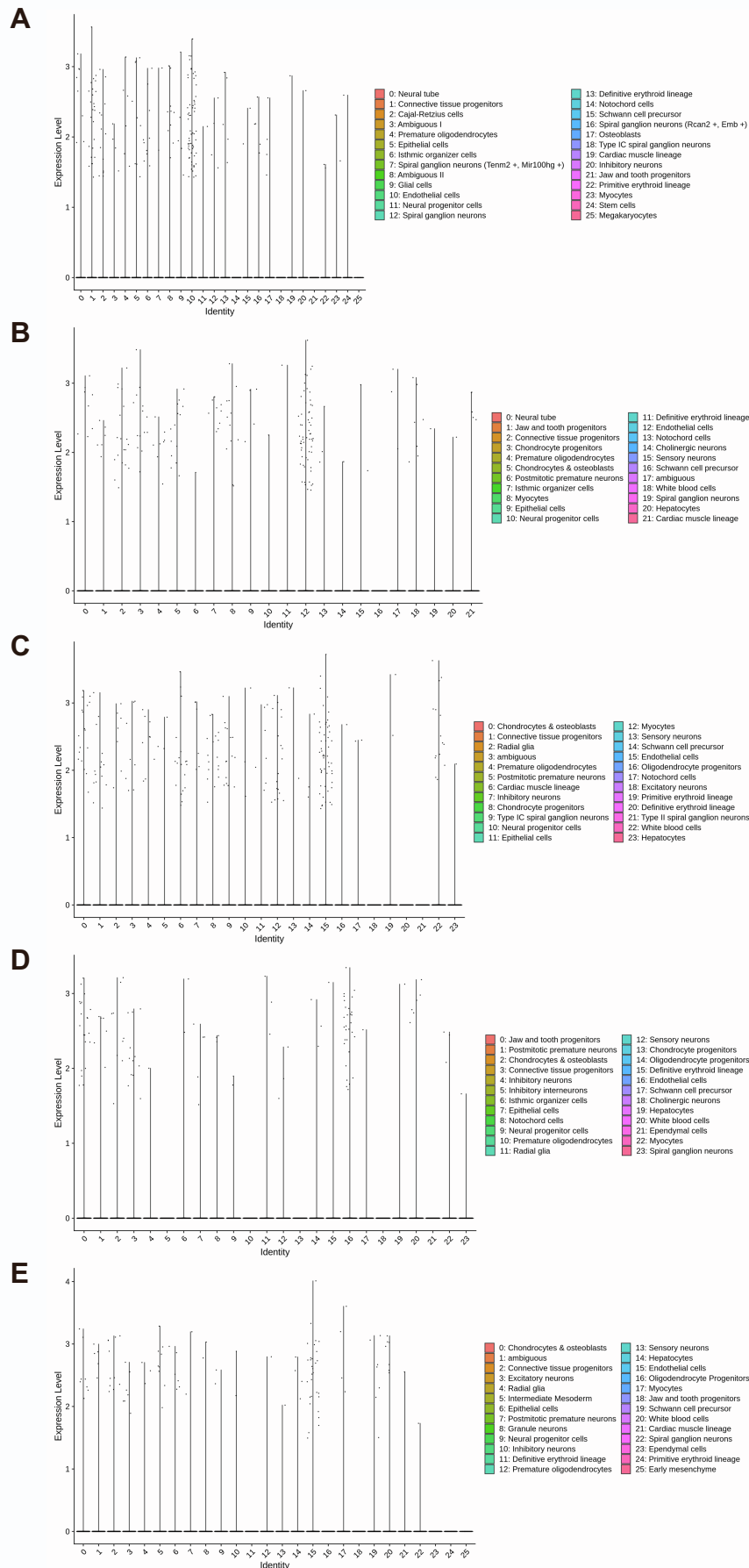


Figure S13. Expression of Activating Transcription Factor 3 in cell clusters from Mouse Organogenesis Cell Atlas at different embryonic days. Analysis of Activating Transcription Factor 3 (*Atf3*) expression on different days from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019): (A) E9.5, (B) E10.5, (C) E11.5, (D) E12.5, and (E) E13.5. The annotation of cell clusters is based on the original 13 publication.

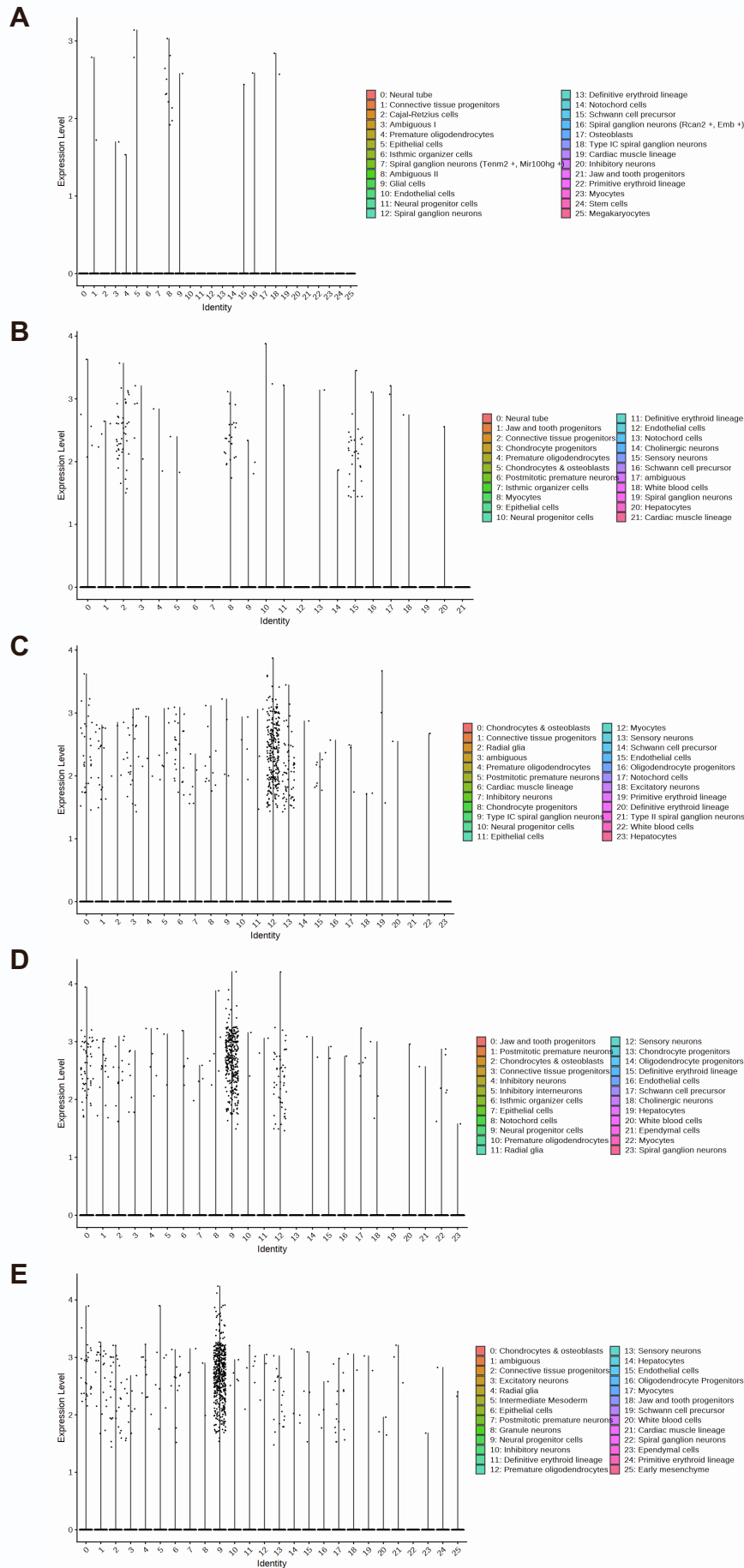


Figure S14. Expression of Musclin in cell clusters from Mouse Organogenesis Cell Atlas at different embryonic days. Analysis of Musclin (*Msc*) expression on different days from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019): (A) E9.5, (B) E10.5, (C) E11.5, (D) E12.5, and (E) E13.5. The annotation of cell clusters is based on the original publication:

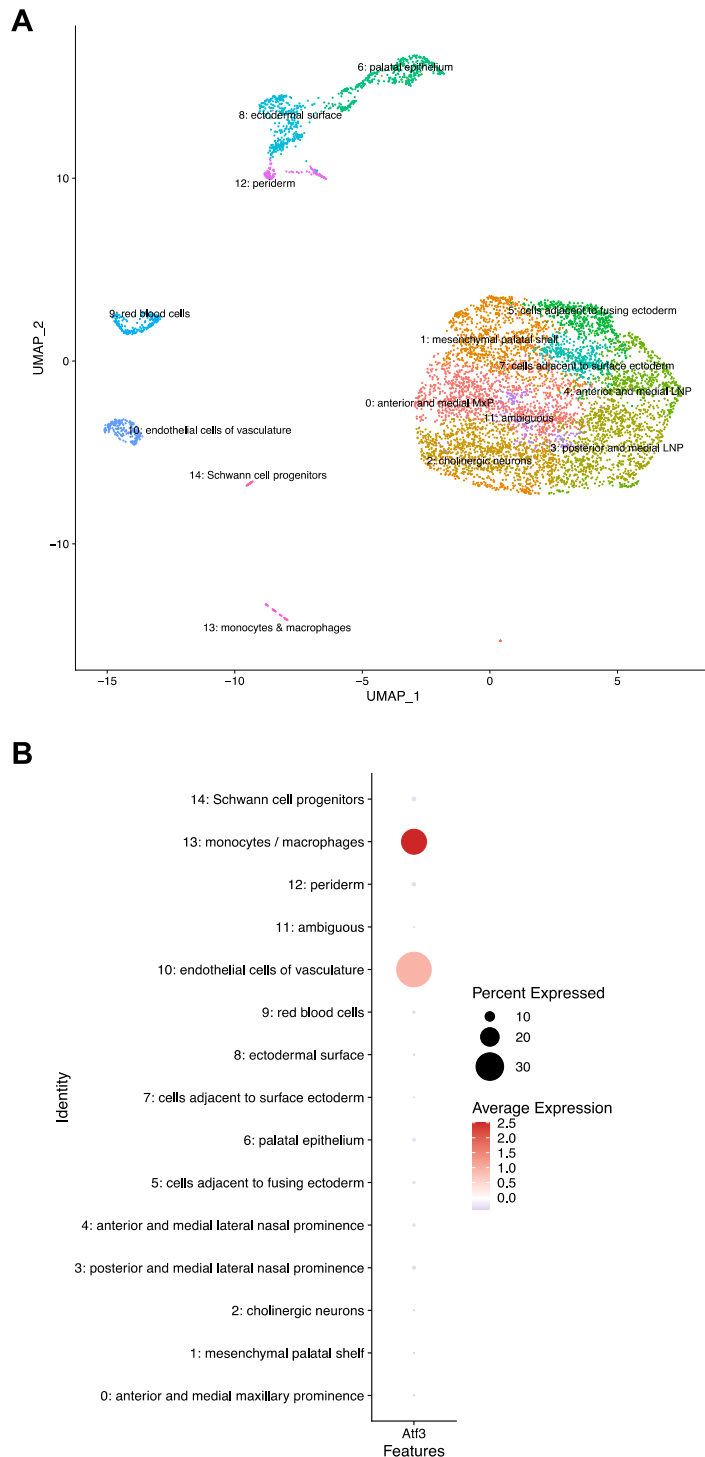


Figure S15. Single-cell expression data of the mouse lambdoidal junction at embryonic day E11.5. (A) Re-analysis of the single-cell data from Li et al. (2019) identified 15 cell clusters that are annotated based on marker gene expression. (B) Single-cell expression data of different cell clusters of the lambdoidal junction at E11.5 are shown as dot plot. For each cell cluster, the percentage of cells expressing *Atf3* is indicated by dot size, while the average expression level is indicated by color. This illustrates, that *Atf3* is mainly expressed in murine monocytes/macrophages and endothelial cells of vasculature. Abbreviation: *Atf3* – Activating Transcription factor 3

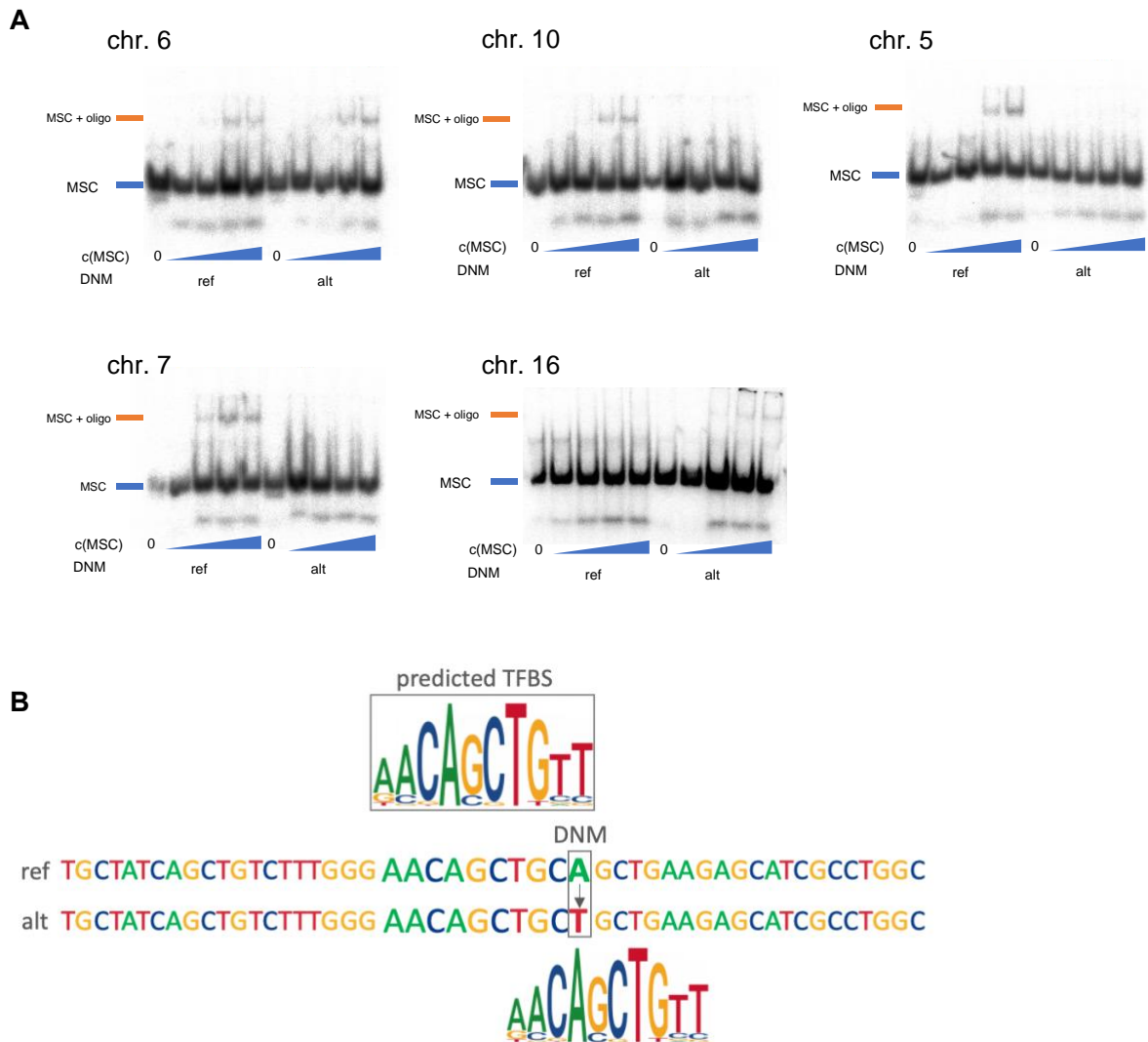


Figure S16. *In vitro* binding of Musculin using Electromobility Shift Assays.

(A) Five genomic regions harboring MSC binding sites and *de novo* mutations (DNMs) affecting the predicted binding affinity were analyzed using EMSA, using oligonucleotides for reference (ref) and alternative (alt) allele. The selected DNMs were: chr6:71445860 G/A; chr10:134303928 G/A; chr5:29647870 A/G; chr7:145175819 A/T; chr16:8870186 C/T. For each candidate binding site, five different concentrations for MSC were titrated for ref (left lanes) and alt (right lanes), respectively. The appearance of the upper band (MSC+oligo) at increasing MSC concentrations reflects a shift in molecular weight, indicating *in vitro* binding of MSC to the oligonucleotide. (B) The predicted transcription factor binding site (TFBS) for the genomic region around the chr7:145175819 A>T DNM from an nsCL/P individual is indicated by the box, with the predicted binding site illustrated above. Upon visual inspection, a second possible binding site was identified that was missed by the *in silico* algorithm, indicated below the DNM with the respective position of the motif. Notably, the new potential TFBS is expected to demonstrate an opposite effect on predicted binding than the original TFBS, which may explain the result of the EMSA experiment.

Abbreviations: MSC – Musculin; EMSA – electrophoretic mobility shift assays; TFBS – transcription factor binding site, DNM – *de novo* mutation; nsCL/P – non-syndromic cleft lip with/without cleft palate

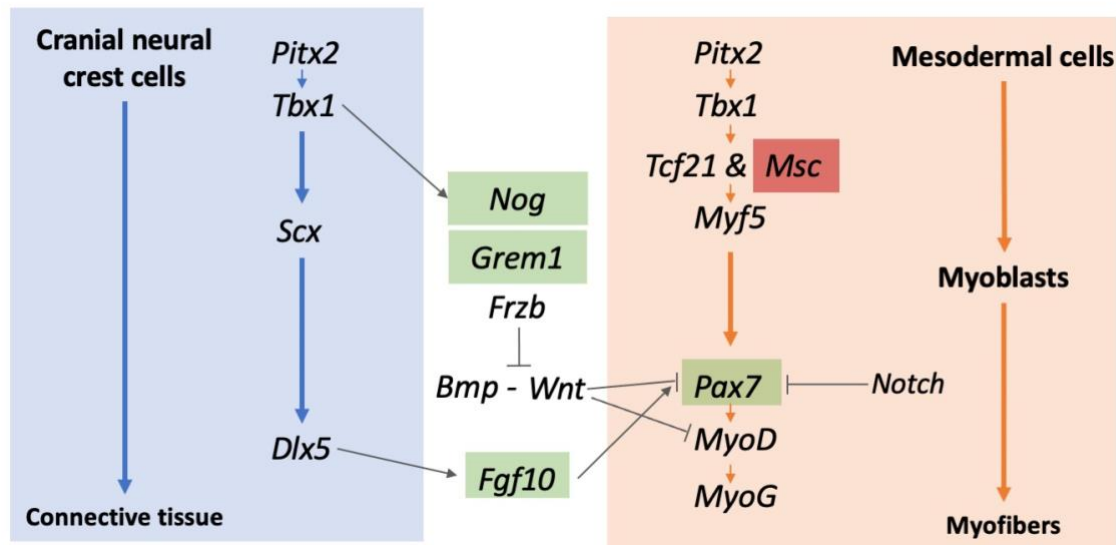


Figure S17. Schematic representation of genes involved in murine embryonic development of branchiomic muscles. The interaction between cranial neural crest cells and mesodermal cells that develop into myofibers via myoblasts is illustrated by orange and blue backgrounds, respectively. Genes identified by genome-wide association studies as candidate genes for non-syndromic cleft lip with/without cleft palate (nsCL/P) are marked in green. Musculin (*Msc*), with binding sites at and binding changes by nsCL/P *de novo* mutations, is highlighted in red. Own illustration based on Salazar et al. (2020).

Tables S1-S3, S5-S6, and S15-S28 are provided as Supplemental Tables in Excel.

See Excel Spreadsheet for Tables S1-S3.

Table S4. Variant effects from Variant Effect Predictor (VEP) and their aggregation in variant effect groups.

Variant effect ^a	Effect names ^b
non	"stop_gained", "stop_gained,splice_region_variant", "start_lost", "stop_gained,NMD_transcript_variant", "start_lost,NMD_transcript_variant", "stop_gained,inframe_deletion"
frame shift	"frameshift_variant", "frameshift_variant,splice_region_variant"
mis	"missense_variant", "missense_variant,splice_region_variant", "missense_variant,NMD_transcript_variant", "missense_variant,splice_region_variant,NMD_transcript_variant", "inframe_insertion", "inframe_deletion", "inframe_insertion,NMD_transcript_variant", "inframe_deletion,NMD_transcript_variant", "missense_variant,splice_region_variant,NMD_transcript_variant"
splice	"splice_donor_variant,NMD_transcript_variant", "splice_donor_variant", "splice_donor_variant,non_coding_transcript_variant"
syn	"synonymous_variant", "splice_region_variant,synonymous_variant", "synonymous_variant,NMD_transcript_variant", "splice_region_variant,synonymous_variant,NMD_transcript_variant"

^a Groups of variant effects: non – nonsense, frameshift, mis – missense, splice – splice site, syn – synonymous.

^b Effect combinations from VEP output for each protein-coding DNM. For annotation of functional effects either Ensembl/Gencode (preferred when available) or RefSeq transcripts were used.

See Excel Spreadsheet for Tables S5-S6.

Table S7. Overview of six *in silico* prediction scores that were used for *de novo* mutation effect prediction.

Annotation Score	Annotation score reference	Threshold ^a
CADD ^b	Kircher et al. 2014	≥ 10 (20)
LINSIGHT	Huang et al. 2017	≥ 0.9
FATHMM	Shihab et al. 2015	≥ 0.9
DANN	Quang et al. 2015	≥ 0.9
ReMM	Smedley et al. 2016	≥ 0.9
ncER	Wells et al. 2019	≥ 95

^a DNMs for which the *in silico* prediction score surpassed the respective threshold were annotated as deleterious.

^b Scaled version of Combined Annotation–Dependent Depletion v1.4. For further prioritization of highly deleterious variants, a more stringent cut-off of 20 was applied.

Abbreviations: LINSIGHT - linear INSIGHT; FATHMM-MKL; DANN - Deleterious annotation of genetic variants using neural networks; REMM - Regulatory Mendelian Mutation; ncER - non-coding Essential Regulation

Table S8. Number of *de novo* mutations included in annotation of different *in silico* prediction scores.

Annotation Score	nsCL/P			NCR			Excluded variants ^a
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic	
CADD	444	6,923	5,387	633	9,181	7,150	1,772
LINSIGHT	246	7,313	5,735	331	9,700	7,596	569
FATHMM	429	6,671	5,204	621	8,843	6,924	2,798
DANN	429	6,701	5,205	622	8,892	6,924	2,717
ReMM	467	7,315	5,740	666	9,703	7,599	0
ncER	461	7,295	5,727	664	9,679	7,589	75

^a Number of *de novo* mutations which been excluded, because no value was output in the respective *in silico* prediction score.

In silico scores are used as described in Table S7. Exonic variants include all variants located in genic regions, including non-coding exons and/or 3'/5' UTRs. The breakdown of variants into bins is reported in Tables S9 to S14. Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S9. Number of *de novo* mutations in different score bins for CADD.

CADD ^a	nsCL/P			NCR		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[1-5[181	5,762	4,401	219	7,643	5,839
[5-10[69	1,116	948	103	1,463	1,227
[10-15[70	305	288	91	445	375
[15-20[45	108	85	76	122	139
[20-30[84	24	17	152	30	17
[30-99]	18	0	1	25	0	2

^a Scaled version of CADD 1.4 (Combined Annotation–Dependent Depletion, Kircher et al., 2014)

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S10. Number of *de novo* mutations in different score bins for ReMM.

ReMM ^a	nsCL/P			NCR		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[0-0.1[91	3,128	2,355	79	4,251	3,112
[0.1-0.2[24	794	462	46	1,093	599
[0.2-0.3[14	580	380	41	797	489
[0.3-0.4[26	627	406	31	834	562
[0.4-0.5[13	655	462	34	834	660
[0.5-0.6[18	539	493	38	676	682
[0.6-0.7[32	441	498	51	541	612
[0.7-0.8[45	297	345	62	341	460
[0.8-0.9[60	153	220	52	202	258
[0.9-1.0]	144	101	119	232	134	165

^a ReMM - Regulatory Mendelian Mutation (Smedley et al., 2016)

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S11. Number of *de novo* mutations in different score bins for FATHMM.

FATHMM ^a	nsCL/P			Non-cleft reference cohort		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[0-0.1[52	3,338	2,188	64	4,542	2,894
[0.1-0.2[108	2,378	2,148	143	3,024	2,850
[0.2-0.3[53	424	445	74	593	584
[0.3-0.4[15	138	112	19	158	146
[0.4-0.5[12	68	52	10	85	71
[0.5-0.6[11	53	39	10	59	51
[0.6-0.7[3	43	31	6	47	37
[0.7-0.8[4	27	28	7	52	53
[0.8-0.9[26	53	37	33	101	65
[0.9-1.0]	145	149	124	255	182	173

^a FATHMM - FATHMM-MKL (Shihab et al., 2015).

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S12. Number of *de novo* mutations in different score bins for DANN.

DANN ^a	nsCL/P			NCR		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[0-0.1[0	23	16	2	42	18
[0.1-0.2[1	198	110	7	224	126
[0.2-0.3[6	358	302	9	571	342
[0.3-0.4[15	689	444	24	928	636
[0.4-0.5[32	999	685	39	1,286	962
[0.5-0.6[42	1,042	865	55	1,452	1,135
[0.6-0.7[61	1,248	1,016	81	1,617	1,327
[0.7-0.8[73	1,240	1,014	83	1,532	1,373
[0.8-0.9[54	688	570	93	950	778
[0.9-1.0]	145	216	183	229	290	227

^a DANN - Deleterious annotation of genetic variants using neural networks (Quang et al., 2015).

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S13. Number of *de novo* mutations in different score bins for LINSIGHT.

LINSIGHT ^a	nsCL/P			NCR		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[0-0.1[162	6,876	5,221	194	9,111	6,888
[0.1-0.2[44	203	313	55	295	386
[0.2-0.3[13	81	59	23	92	92
[0.3-0.4[8	43	20	10	37	49
[0.4-0.5[4	26	25	9	33	44
[0.5-0.6[0	11	16	7	18	17
[0.6-0.7[2	8	10	1	20	12
[0.7-0.8[0	9	7	4	7	12
[0.8-0.9[7	21	22	12	42	36
[0.9-1.0]	6	35	42	16	45	60

^a LINSIGHT - linear INSIGHT (Huang et al., 2017).

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

Table S14. Number of *de novo* mutations in different score bins for ncER.

ncER ^a	nsCL/P			NCR		
	Exonic	Intergenic	Intronic	Exonic	Intergenic	Intronic
[0-50[66	4,153	2,986	88	5,570	3,942
[50-80[71	2,330	1,218	81	3,074	1,615
[80-90[47	508	738	73	649	978
[90-95[80	164	409	111	211	567
[95-99[135	123	322	219	156	385
[99-100]	62	17	54	92	19	102

^a ncER - non-coding essential regulation (Wells et al., 2019).

Abbreviations: nsCL/P - non-syndromic cleft lip with/without cleft palate; NCR - non-cleft reference cohort

See Excel Spreadsheet for Tables S15-S28.

Table S29. Transcription factors with a significant excess of hits and change of binding for nsCL/P *de novo* mutations, compared to those in NCR.

Motif name	Qualitative analysis of number of hits			Quantitative analysis of binding change		
	Ratio (nsCL/P:NCR) ^a	Log2FC ^b	P-value ^c	Ratio (nsCL/P:NCR) ^d	Log2FC ^e	P-value ^f
JDP2 (var.2)	3.34 (5:2)	1.74	0.1256	2.32	1.21	-
MSC	4.68 (7:2)	2.23	0.0371	2.42	1.28	-
MEF2A	2.01 (6:4)	1.00	0.2163	4.07	2.03	<i>0.025</i>
MAF::NFE2	2.68 (2:1)	1.42	0.3923	8.19	3.03	-
ATF3	4.68 (7:2)	2.23	0.0371	2.93	1.55	-
SRF	2.68 (2:1)	1.42	0.3923	3.09	1.63	-
NFE2L1	2.68 (2:1)	1.42	0.3923	60.49	5.92	-

^a Ratio of nsCL/P and NCR DNMs with hits by specific position weight matrix (PWM) of transcription factor, corrected for total number of hits per cohort. Absolut number of hits in both cohorts in brackets (nsCL/P vs. NCR cohort).

^b Log2FC of DNM ratio per PWM, corrected for total number of hits per cohort.

^c Fisher's Exact Test; Motifs with nominally significant findings are represented in bold.

^d Ratio of mean binding change by DNM for the respective PWM between nsCL/P and NCR DNMs.

^e Log2FC of ratio of mean binding change between cohorts.

^f Mann-Whitney-U-Test (MWU-Test), nominally significant findings in italic; "–" indicates that no MWU-Test was performed. Motifs were excluded from MWU-Testing if there was: (i) less than 3 DNM-PWM hits per cohort; and/or (ii) lack of variability in change of binding (exclusion of 168 motifs in total).

Abbreviations: DNM – *de novo* mutation; nsCL/P – non-syndromic cleft lip with or without cleft palate; NCR – non-cleft control cohort; Log2FC – Log2 Fold Change;

Table S30. Summary of results of electromobility shift assays for binding change of Musculin to oligonucleotides carrying DNM reference or alternative allele.

Position	Cohort	Predicted binding change ^a	Replicate 1 ^b	Replicate 2	Replicate 3
chr6:71445860 G/A	nsCL/P	Loss (-8.43)	Small gain	No change	Small gain
chr7:145175819 A/T	nsCL/P	Gain (+8.43)	Loss	Loss	Loss
chr10:134303928 G/A	nsCL/P	Loss (-8.43)	Loss	Loss	Loss
chr16:8870186 C/T	nsCL/P	Gain (+8.43)	Gain	Gain	Gain
chr5:29647870 A/G ^c	NCR	Loss (-4.44)	Loss	Loss	Loss

For each candidate binding site, electro mobility shift assay (EMSA) was performed in triplicates.

^a Predicted binding change of the transcription factor Musculin to genomic sequence around the respective DNM using the position weight matrix from JASPAR 2020 in a modified version of denovoLOGOB (prediction of binding change can be categorized into gain of binding (if PWM-ref<PWM-alt), loss of binding (PWM-ref>PWM-alt), and silent effects (PWM-ref=PWM-alt))

^b Representative figures of EMSA are shown in Figure S16A (Replicate 1)

^c Binding site at + strand, original DNM base exchange: T/C

Abbreviations: DNM – *de novo* mutation; nsCL/P – non-syndromic cleft lip with or without cleft palate; NCR – non-cleft control cohort

Table S31. *De novo* mutations in non-syndromic cleft lip with/without cleft palate cohort in *ZFHX4*.

DNM ^a	REF ^b	ALT ^b	CADD	ReMM	DANN	FATHMM	LINSIGHT	ncER
Chr8:77621099	T	A	13.76	0.697	0.753	0.593	0.254	97.84
Chr8:77647464	G	A	1.22	0.453	0.267	0.188	0.068	88.11
Chr8:77764751	CA	C	-	0.938	-	-	-	99.66

- indicates that there is no value for this variant for the specific *in silico* score. Scores highlighted in bold represent scores surpassing the respective threshold as shown in Table S7. All abbreviations and references provided in Table S7.

^a DNM position according to genome assembly version hg19 (GRCh37).

^b REF shows reference allele at genomic position in hg19, ALT represents observed DNM.

Abbreviations: DNM – *de novo* mutation; nsCL/P – non-syndromic cleft lip with or without cleft palate; REF – reference allele,

Supplemental Methods

Datasets and variant calling

Whole genome sequencing (WGS) data were previously generated as part of the Gabriella Miller Kids First (GMKF) project. For non-syndromic cleft lip with/without cleft palate (nsCL/P), data was generated by the *Genomic Studies of Orofacial Clefts Birth Defects* and was accessed through dbGaP upon approved data access (phs001168.v1.p1). The raw sequencing WGS dataset included 1,236 individuals from case-parent trios with different types of orofacial clefts (OFC). Phenotypic information included: subject IDs, father and mother IDs, sex, ethnicity, race, cleft type, and evidence of non-isolated cleft. Based on phenotypic information, we excluded trios with (i) missing WGS data for one of the three family members (n= 80 trios), (ii) affected parent(s) (n= 42 trios), and (iii) any other type of OFC than nsCL/P (n= 70 trios). The final pre-variant calling dataset comprised 220 nsCL/P trios. This study cohort represents a subcohort of a previously published study on coding *de novo* mutations by Bishop et al.¹ For the non-cleft reference (NCR) cohort, we retrieved WGS data from 330 case-parent trios from the “Genetic Contribution to Ewing Sarcoma” cohort (access through dbGaP, accession number: phs001228.v1.p1). This cohort comprises primarily individuals of predominantly European descent (according to PubMed ID: 35512711) and has already been used as validation cohort in two Ewing Sarcoma studies.^{2,3} After filtering for trio completeness the dataset comprised 289 trios. Alignment of fastq-data and subsequent variant calling was performed as previously described⁴ and equally applied to both cohorts. Briefly, reads were aligned to the reference genome GRCh37 using bwa-mem. Subsequently, single-nucleotide variants (SNVs) and small indels were called using UnifiedGenotyper (after realignment) and HaplotypeCaller tools from Genome Analysis Tool Kit v3.7, with default settings.⁵ Next, probable *de novo* mutations (DNMs) were identified (defined as heterozygous genotype in the index patient and homozygous genotype in the parents). For the present study, variant identification was restricted to autosomal DNMs. We further refined our dataset by excluding case-parent trios with DNMs above median + the 3. IQR (9 nsCL/P trios and 5 NCR trios excluded), and only retained variants with quality score >140 and MQRank Sum > -5.5, resulting in a final dataset of 211 nsCL/P and 284 NCR trios. Cut-offs were determined based on the combined datasets using histograms (Figure S2-S4). The distribution of cleft phenotype and sex of the final 211 nsCL/P cases are shown in Figures S5-S6.

Additional consideration for using trios with Ewing Sarcoma phenotype as controls

There is epidemiological evidence for some shared genetic influences on cancer and facial clefting. However, so far and to our knowledge, these have not been confirmed at molecular level. Given the general paucity of publicly available WGS trio data, the Ewing Sarcoma (ES) cohort was chosen as it was highly matching the nsCL/P case cohort from a study design perspective: (i) it included predominantly European individuals, (ii) there is only limited evidence for a role of germline mutations in ES, and (iii) data were generated using the similar platforms (i.e., HiSeq X) and harmonized pipelines in the GMKF project. Therefore, any artifacts and biases related to those parts of our analyses could be excluded. Still, ES patients are not considered population-based or healthy individuals as theoretically, they may also have some accumulation of DNMs as part of the disease etiology of ES (although this is not yet reported). Such (yet unknown) effect would result in limited power for our study, as we might miss DNMs (or a regional enrichment thereof) at loci that play a role in

both disorders. Hence, the selection of this cohort might result in false negatives due to limited power but does not impose the risk of false positives.

DNM annotation

All DNMs. DNMs were classified as intronic, exonic, or intergenic based on positional information and the GENCODE Basic gene annotation version33.hg19 (downloaded in February 2020).⁶ The list of transcripts (n=20,084) was filtered for protein-coding genes and autosomal location, leaving 19,145 protein-coding genes for analysis. In case a DNM mapped to multiple transcripts, exonic positions were preferred over intronic positions. Exonic DNMs hereby included all variants located in genic regions, including non-coding exons and/or 3'/5' UTRs. All DNMs that could not be mapped to exonic or intronic regions of this gene set were classified as intergenic.

All DNMs were annotated with information on frequency (gnomAD v3.1, all populations; Figure S1). No general allele frequency filter was applied to dataset.

For each DNM we retrieved six different *in silico* prediction scores from respective databases, i.e., CADD (Combined Annotation-Dependent Depletion),⁷ ReMM (Regulatory Mendelian Mutation),⁸ FATHMM (Functional Analysis through Hidden Markov Models),⁹ DANN (Deleterious annotation of genetic variants using neural networks),¹⁰ LINSIGHT (linear INSIGHT)¹¹, and ncER (non-coding Essential Regulation).¹² Applied thresholds are listed in Table S7.

Subset of protein-coding DNMs. For each protein-coding DNM, the Ensembl Variant Effect Predictor (VEP, see Web Resources) tool¹³ was used to annotate functional effects using either Ensembl/GENCODE (preferred when available) or RefSeq transcripts. Analysis was limited to five groups (nonsense, frameshift, missense, splice, and synonymous; Table S4). In case of multiple assignments for a DNM, we prioritized these effects according to effect strength (nonsense>frameshift>missense>splice>synonymous). We also grouped these DNMs further into Loss of function (LoF; includes nonsense, frameshift, and splice effects) and protein-altering DNMs (LoF and missense).

Comparison of exonic DNMs with DNMs identified by Bishop et al. (2020)

As the nsCL/P cohort from GMKF in our study represents a subcohort of Bishop et al., this allowed us to compare coding DNMs between both studies for variant calling control, at least for those individuals. For this comparison, we used our entire set of genome-wide DNMs and all 862 rare coding DNMs identified by Bishop et al. (2020)¹. As Bishop et al. had included DNMs from trios of different ethnicities, we restricted the Bishop et al. variants to those observed in Europeans and in patients with phenotypes 2 (CLO) and 3 (CLP; Table S3 in Bishop et al., 2020)¹. This resulted in 323 DNMs from 206 different samples, whose coordinates were then transferred to hg19 (GRCh37) for comparison. Based on variant position in Bishop et al. we identified sample IDs and DNM overlaps, and also analyzed our pre-QC dataset for variants that were absent from our study but observed in Bishop et al. Together the results indicate that variants exclusive to one study are attributed to QC parameters in the individual studies. We provide a summary table with all coding DNMs from both studies, including their sample overlap, in Table S6.

Statistical comparison of DNM distribution between cohorts

The average number of DNMs per sample was compared between cohorts using a Mann-Whitney-U-Test. Analysis was performed for all DNMs, and for the subgroups of exonic,

intronic, and intergenic DNMs. The distribution of *in silico* prediction scores for nsCL/P and NCR DNMs was compared by the percentage distribution of the score values for the entire dataset of DNMs in nsCL/P and NCR cohort and for the subset of non-coding DNMs (Figure 1B, Figure S8).

For raw CADD scores, a similar distribution between cohorts was shown before using scaled CADD scores for all analyses (Figure S9).

To compare the proportion of DNMs with particularly high *in silico* prediction scores among cohorts, chi-squared tests were used for the number of DNMs over the respective threshold compared to the rest of DNMs with lower scores (Thresholds in Table S7). For DNMs exceeding the threshold in 5 or 6 respective *in silico* scores, we tested the number of DNMs above the appropriate number of thresholds against variants that did not meet the respective thresholds. The number of DNMs exceeding the threshold of multiple *in silico* prediction scores was also compared by the percentage distribution (Figure S10).

Additionally, the number of DNMs with scaled CADD score ≥ 20 (i.e, top 1% of ranked reference genome SNVs), were compared to those DNMs with CADD < 20 as a more stringent cut-off.

Statistical enrichment analysis

For calculating enrichment in different sets of functional elements, the R package FunciVar was used.¹⁴ In FunciVar, enrichment analyses are based on a Bayesian version of the binomial test (for details see Jones et al., 2020).¹⁴ Briefly, FunciVar simulates a distribution of enrichment probabilities for two sets of variants (10,000 simulations by default). Then, the distribution of differences between the two enrichment probabilities is computed and, finally, a 95% credible interval for the range of enrichment probability differences between the two lists of variants is determined. In FunciVar, the significance of the results is given as the probability (data range: 0 to 1) that variants in the candidate set group (here: nsCL/P DNMs) have more overlap with the dataset of functional elements than variants in the background group (here: NCR DNMs). The closer the probability value to 1, the more likely is a significant difference between the two sets of variants. FunciVar returns the 95% credible interval of the difference of enrichment probabilities and the median of the credible interval as point estimate of enrichment (ranges between -1 to 1, with 1 meaning strong enrichment, and -1 meaning strong depletion) along with the probability of enrichment. To bring the Bayesian approach closer to the frequentist interpretation of the remaining results of this paper, we calculated a p-value equivalence based on the probability of direction (P_d)¹⁵. The approximate p-value is calculated as $P=2*(1-P_d)$, which corresponds to the approximate relationship between the frequentist p-value and the P_d .¹⁶ However, it must be emphasized that the P_d actually has a different interpretation than the frequentist p-value and this p-value conversion is intended only for a simpler interpretation, in line with the remaining results of this article to readers non-familiar with Bayesian statistics. The detailed results of the Bayesian approach (the effect estimator, the credibility intervals, and the probability of enrichment) are presented in Supplemental Tables (Tables S17, S20, S22-25).

For each enrichment analysis, candidate and background groups were defined as set of variants located within and outside of the tested elements (see Datasets used for enrichment analyses).

Datasets used for enrichment analyses

Chromatin data of facial development. As regulatory effects are cell-type and time-point specific, we retrieved epigenetic datasets that were drawn from cell types and

developmental stages of relevance for facial development, namely: (i) *in vitro* chromatin states in early human neural crest cells (hNCC)¹⁷ and cranial neural crest cells (cNCC)¹⁸ (GEO; hNCC: GSE28874, cNCC: GSE70751), and (ii) chromatin states generated in human craniofacial tissue (CT) of multiple time points in craniofacial development (Carnegie stage (CS) 13, CS14, CS15, CS17, CS20, 10 weeks *post conceptionem*); GSE97752).¹⁹ Joint data processing using an in-house pipeline has been previously performed²⁰, and data is available at Zenodo (doi: 10.5281/zenodo.3911187). The final output of this analysis were chromatin states corresponding to eight states: transcription start site (TSS), transcription (Tx), enhancers (Enh), ZNF genes and repeats (ZNF_Rpts), Heterochromatin (Het), bivalent/poised transcription start site or bivalent enhancer (TssBiv_EnhBiv), repressed Polycomb (ReprPC), and Quiescent/Low (Quies). For each state and tissue/cell type, enrichment of nsCL/P DNMs was calculated using FunciVar as described above, resulting in 64 tests for DNM enrichment analysis by chromatin state data. Benjamini-Hochberg procedure was used for the correction for 64 tests.

Conserved regions. Based on the hypothesis that highly conserved non-coding elements could be relevant for conserved facial development, we retrieved a dataset of 4,307 evolutionarily highly conserved non-coding elements (CNEs, see Data and Code availability) from a prior study of DNMs in regulatory elements in neurodevelopmental disorders.²¹ These CNEs were tested for enrichment of nsCL/P DNMs using FunciVar, as described above.

VISTA enhancer. We retrieved 2,974 *in vivo* tested elements with tissue-specific enhancer activity from the VISTA database²² (see Web Resources, accessed 2019/10/24). Of those, 1,570 showed enhancer activity, were located on autosomes, and could be unambiguously mapped to the human genome (only mice enhancers (genome mm9) with associated hg19 coordinates of human enhancer in VISTA Enhancer Browser were included). Enhancer activity in VISTA is defined as a reproducible expression in the same structure in at least three independent transgenic embryos. First, all those VISTA enhancers were tested for enrichment of nsCL/P DNMs with FunciVar. Subsequently, VISTA enhancers were grouped based on tissue-specific enhancer activity, in order to identify DNM enrichment in specific tissue-related enhancers. Therefore, enrichment was calculated for every enhancer group that was reported active in a specific tissue (using the same criteria for activity, i.e., reproducible expression in this structure in at least three independent transgenic embryos) and also contained nsCL/P and/or NCR DNMs. This resulted in 16 tests (i.e., 16 tissues showed active enhancers in which DNMs from our dataset were localized, Table S23). For DNMs mapping in regions with multiple overlapping enhancers, DNMs were considered active for all tissues with activity of enhancers (total: n=4: one DNM in two human enhancers: chr13:95618516-95619850, chr13:95618464-95619819; 3 DNMs (2 nsCL/P, 1 NCR) within overlapping human and mouse enhancers (2 nsCL/P DNMs within chr1:181121049-181123654, chr1:181118450-181122869, 1 NCR DNM within chr10:134442029-134446812, chr10:13444023-134446722).

Topologically associating domains (TADs). Based on a dataset of 2,991 autosomal TADs from human embryonic stem cells (hESC),²³ DNMs were mapped to these regions based on positional location. We also defined a subset of 45 TADs that encompassed common risk variants from 45 GWAS loci identified in previous studies (TADs_{GWAS}, Table S1, based on Welzenbach et al., 2021²⁰). For locus 8q22.1 surrounding TADs (centromeric TAD: 8q22.1(I), telomeric TAD: 8q22.1(II)) were tested for enrichment of nsCL/P DNMs. Two

GWAS loci (5p12, Yu et al., 2017;²⁴ Welzenbach et al., 2021²⁰) were mapped into one TAD. Enrichment analyses with FunciVar were performed for all TADs, and for the subset of TADs_{GWAS}. Benjamini-Hochberg procedure was used for correction for multiple testing for 2,961 (i.e., those TADs in which DNMs were present) and 45 tests, respectively.

Analysis of transcription factor binding sites (TFBS)

Mapping of DNMs to PWMs. For each DNM, we analyzed potential effects of its reference (ref) and alternative (alt) allele on transcription factor (TF) binding sites (TFBS). To predict and quantify changes in TF binding for each DNM at a potential TFBS, we used a modified version of the tool denovoLOGOB (short for *de novo Loss of Binding/Gain of Binding*, previously denoted as denovoTF, available on GitHub (see Web Resources)).²¹

DenovoLOGOB predicts TF binding to a genomic region around an SNV by analyzing the consistency of genomic sequences around ref and alt allele with position weight matrices (PWMs) of TF binding motifs. Changes in the denovoLOGOB package included (i) the integration of JASPAR 2020, (ii) the evaluation of binding events for both genomic strands with separate scripts (core_plus and core_minus), and (iii) the default calculation for TF binding with alt allele when binding was present in ref allele (above limit value of 95%) to ensure that the maximum binding change (BC) was detected for every PWM-DNM combination. All scripts used are available at Zenodo: 10.5281/zenodo.5601707.

To retrieve PWMs for human TFs, the Bioconductor package JASPAR2020²⁵ (see Web Resources) with 810 PWM was integrated in denovoLOGOB using TFBSTools²⁶ (see Web Resources). Values for TF binding at DNM positions were calculated for all possible positions of DNMs within each PWM (genomic sequence length: DNM +/- motif length - 1). A high value for a PWM at a DNM position indicates a potential stronger binding of the TF to the genomic region: to filter for such sufficient binding sites, only TFBS where the genomic region of the DNM for ref or alt allele reaches a threshold value of 95% of the potential value range of the PWM are displayed as possible binding sites in denovoLOGOB output (≥ 95 . quantile between minimal and maximal possible value from PWM).

Statistical analysis. Comparing the consistency of the genomic sequence for ref and alt with PWMs for the selected DNM-PWM combinations reveals the BC effect by DNM, which can be categorized into gain of binding (if $\text{PWM-ref} < \text{PWM-alt}$), loss of binding ($\text{PWM-ref} > \text{PWM-alt}$), and silent effects ($\text{PWM-ref} = \text{PWM-alt}$). This value of BC between ref and alt allele is reported as absolute value of difference in PWM-DNM consistency.

The analysis of TFBS with denovoLOGOB was limited to SNVs (deletions and insertions were omitted). In case that multiple DNM-PWM hits per DNM-PWM were observed, we prioritized DNM-PWM combinations with highest absolute BC. In case of same change of binding for + and - strand, we preferred + over - strand. Number of DNM-PWM hits between cohorts was compared using a chi-squared test using number of all possible DNM-PWM combination with JASPAR 2020 dataset of 810 PWMs for the included 12,335 nsCL/P DNMs vs. 16,438 NCR DNMs (9,991,350 vs. 13,314,780 possible PWM-DNM hits for nsCL/P and NCR cohort, respectively).

Joint analysis of DNMs for individual PWMs. DNM-PWMs were grouped based on PWM identity, i.e., by their overlapping binding motif. For a single TF, multiple motifs are available in the JASPAR database. Note: since these binding motifs and their corresponding PWM

differ greatly in some cases, we did not collapse these motifs according to their TFs, but treated them separately. We then statistically analyzed the number of absolute hits per cohort and assessed the quantitative changes in the binding strength (as absolute BC). First, to identify a significant excess of nsCL/P DNM hits for individual PWMs, a Fisher's Exact test was performed for all PWMs in combination with a log2fold change (log2FC) of binding events in cohorts ($\log_2\text{FC} = \text{nsCL/P hits}_{\text{corr}} \left(\frac{\text{hits(PWM)}}{\text{all hits}} \right) / \text{NCR hits}_{\text{corr}} \left(\frac{\text{hits(PWM)}}{\text{all hits}} \right)$).

For the calculation of log2FC, those PWMs with hits in only one cohort were excluded. Analysis of quantitative BC was performed using absolute values of the binding difference between ref and alt allele. The Mann-Whitney-U (MWU) -Test was performed for all PWMs with at least 3 DNM-PWM hits per cohort and variability in BC (exclusion of 168 motifs with less than 3 hits in at least one cohort, or missing variance in cohorts).

Log2FC was calculated using the ratio of mean binding change from hits in nsCL/P and NCR for each PWM.

To extract PWMs with more binding hits and higher BC by nsCL/P DNMs, we selected all PWMs with log2FC of hits ≥ 1 , and a log2FC ≥ 1 BC. We then filtered these TFs (PWMs), for PWMs that had either a significant MWU-Test or a significant Fisher's exact test. For PWMs, for which an MWU-Test was not computable an additional filter was applied for integration in box of Figure 3B with MWU-Test results: total number of hits ≥ 5 . PWMs that met the defined criteria (log2FC ≥ 1 for both approaches and one of the following criteria: significant MWU-Test/Fisher's Exact Test/MWU-test missing) were defined as the overlap of approaches and are shown in Table S29.

Single-cell expression analysis in mouse embryonic development

We used recently generated single-cell expression data from whole mouse embryos (Mouse Organogenesis Cell Atlas, MOCA)²⁷ as well as the lambdoidal junction,²⁸ to analyze the expression of candidate TFs in cell types involved in nsCL/P development.

Re-analysis of MOCA. The MOCA dataset (Processed/Sampled/Split Data/gene_count_cleaned.RDS under

<https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/downloads>) comprised over 1.3 million filtered high quality cells from E9.5 to E13.5, and was split into 5 different datasets, i.e., one per embryonic day (112,269 cells at E9.5, 258,104 cells at E10.5, 449,614 cells at E11.5, 270,197 cells at E12.5 and 241,800 cells at E13.5). For single-cell based gene analysis, the R toolkit Seurat v4.0²⁹ was used. Data was first log normalized using default settings, then scaled, and subsequent feature selection was performed by choosing 2,500 highly variable genes using the vst selection method. Principal component analysis was computed on the variable features from feature selection. For clustering, first the k-nearest neighbors of each cell were identified based on the first 25 principal components. Then, the modularity was optimized using the Louvain algorithm at a resolution of 0.5. Marker genes for each cluster were calculated with the Wilcoxon Rank Sum test using only positive markers and a minimum fraction of 0.25 of cells expressing the respective gene in either of the tested populations. The annotation of cell types was performed using the marker genes published by Cao et al. in 2019 and the R package scCATCH.³⁰ UMAP was based on the first 25 principal components (Figure S12).

Re-analysis of Li et al. The single-cell dataset of the lambdoidal junction from the murine face at E11.5 (GEO; GSM3867275) contained a post-filtering set of 7,249 high quality cells. Filtering for high quality cells included: (i) 2,300-7,500 unique genes to exclude apoptotic or

lysed cells as well as doublets, (ii) a number of RNA counts < 80,000 to exclude doublets, and (iii) a percentage of less than 5% of mitochondrial genes per cell to exclude lysed cells. Data was log normalized using default settings, scaled and feature selection was performed the same way as for MOCA. Principal component analysis was computed on the variable features from feature selection. Cell clustering was performed the same way as for MOCA. The annotation of cell types was performed using the marker genes published by Li et al., 2019.²⁸ UMAP was based on the first 25 principal components (Figure S15A).

Electrophoretic mobility shift assays

To study DNA-protein interaction via electrophoretic mobility shift assays (EMSA), pET-28a(+) harboring MSC with C-terminal His6-tag was ordered from ATG:biosynthetics. The plasmid was transformed into *E. coli* BL21 (DE3). Cells were cultivated overnight at 37°C and 180 rpm until OD=0.6. Expression of *Musculin* was induced by using 1mM IPTG and incubated for 5 h [KL1] at 37°C and 166 rpm. The cells were collected, frozen on dry ice and stored at -80 °C.

Preparation of cleared lysates and purification of Musculin. Cell pellets were thawed and resuspended in lysis buffer (5ml/1g cell pellet, 50mM NaH₂PO₄, 300mM NaCl, 10mM imidazole, pH=8.0). Lysozyme was added, and lysate was incubated on ice for 30 min, followed by sonication on ice (3min with pulse 10sec on, 5sec off). After centrifugation (10,000 x g for 30 min at 4°C), cleared lysate was added to Ni-NTA (Qiagen, ratio 2:1). After incubation for 1h at 4°C while shaking, lysate-Ni-NTA mixture was washed once with 4 column volumes of lysis buffer (50mM NaH₂PO₄, 300mM NaCl, 10mM imidazole, pH=8.0), twice with four column volumes of wash buffer (50mM NaH₂PO₄, 300mM NaCl, 20mM imidazole, pH=8.0). For elution 4x elution buffer (0.5ml) was added and samples were collected (50mM NaH₂PO₄, 300mM NaCl, 250mM imidazole, pH=8.0). Eluates were separated on 12% SDS-PAGE, visualized by Coomassie Blue Staining and Western Blot analysis against anti-His. Protein concentration was determined by using photometric measurements.

EMSA. Oligonucleotides of five DNA sequences with harboring a DNM that is predicted to affect Musculin binding were ordered from Sigma (binding motif at DNM position +/- 20 bp with ref and alt allele for DNM position). After dissolving lyophilized oligonucleotides in water, oligonucleotides were annealed to the reverse complement strand to achieve double stranded DNA. All DNA-binding reactions were performed in 1x binding buffer (10mM Tris (pH 7.5), 100mM NaCl, 1mM MgCl₂, 10% Glycerol). 10nM DNA was incubated with five different concentrations of Musculin (range 0-1μM). The reactions were incubated for 15 min at 21°C and then loaded on an 8% native polyacrylamide gel (19:1) in 1x TBE buffer. The electrophoresis was performed using constant 6V/cm. The gels were vacuum-dried, exposed to a phosphor screen, and visualized using a Typhoon Phosphoimager. For each tested DNM-binding reaction, three replicates were performed for reference and alternative alleles.

Supplemental References

1. Bishop, M.R., Diaz Perez, K.K., Sun, M., Ho, S., Chopra, P., Mukhopadhyay, N., Hetmanski, J.B., Taub, M.A., Moreno-Urbe, 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.
2. Miller, D.B., and Piccolo, S.R. (2021). A Survey of Compound Heterozygous Variants in Pediatric Cancers and Structural Birth Defects. *Front. Genet.* *12*, 363.
3. Gillani, R., Camp, S.Y., Han, S., Jones, J.K., Chu, H., O'Brien, S., Young, E.L., Hayes, L., Mitchell, G., Fowler, T., et al. (2022). Germline predisposition to pediatric Ewing sarcoma is characterized by inherited pathogenic variants in DNA damage repair genes. *Am. J. Hum. Genet.* *109*, 1026–1037.
4. Holtgrewe, M., Knaus, A., Hildebrand, G., Pantel, J.T., Santos, M.R. de los, Neveling, K., Goldmann, J., Schubach, M., Jäger, M., Coutelier, M., et al. (2018). Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. *Sci. Rep.* *8*, 14611.
5. Depristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* *43*, 491–501.
6. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* *47*, D766–D773.
7. 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.
8. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurphy, 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.
9. 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.
10. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* *31*, 761–763.
11. 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.
12. 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.
13. 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.
14. 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.
15. Makowski, D., Ben-Shachar, M.S., Chen, S.H.A., and Lüdtke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Front. Psychol.* *10*, 2767.

16. Makowski, D., Ben-Shachar, M.S., and Lüdtke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *J. Open Source Softw.* 4, 1541.
17. 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.
18. 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 Chimpanzee Neural Crest. *Cell* 163, 68–83.
19. Wilderman, A., VanOudenhoove, J., Kron, J., Noonan, J.P., and Cotney, J. (2018). High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell Rep.* 23, 1581–1597.
20. 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. *Hum. Genet. Genomics Adv.* 2, 100038.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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. (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48, D87–D92.
26. Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 32, 1555–1556.
27. 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.
28. 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.* 146, dev174888.
29. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29.
30. Shao, X., Liao, J., Lu, X., Xue, R., Ai, N., and Fan, X. (2020). scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *IScience* 23, 100882.