

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>Aff3</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

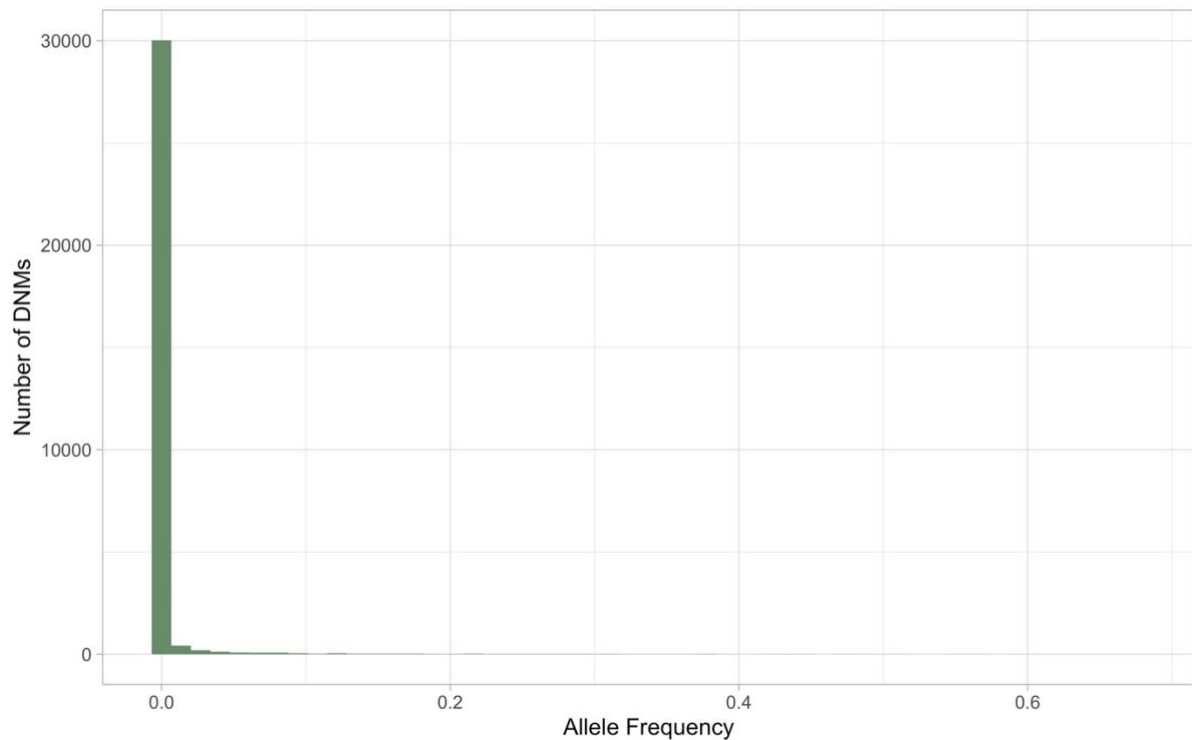


Figure S1. Distribution of allele frequency of all *de novo* mutations in dataset.

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

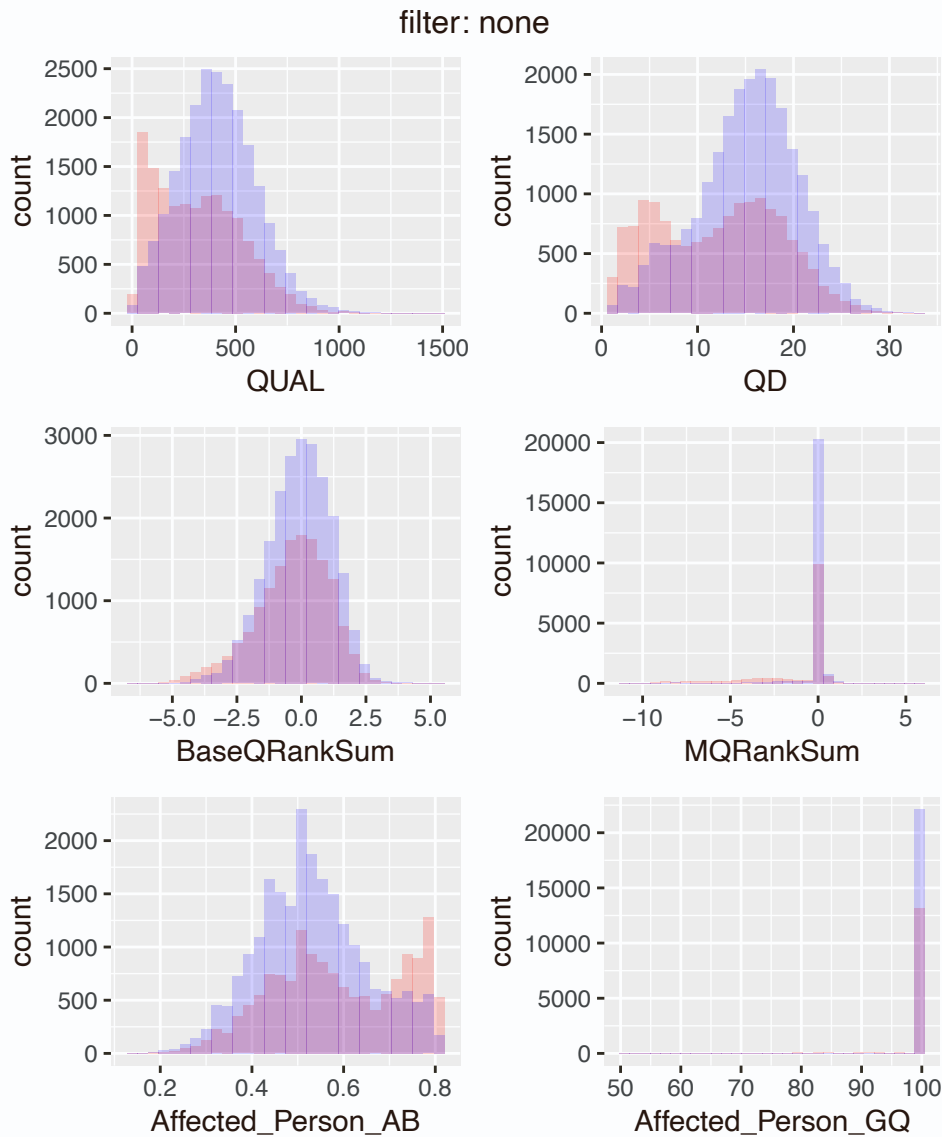


Figure S2. Different quality parameters of *de novo* mutations. The individual histograms show quality scores of *de novo* mutations (post sample QC; intersect between Haplotype Caller and Unified Genotyper). The QUAL, QD, BaseQRankSum, and Affected_Person_GQ values are the values determined for the 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.

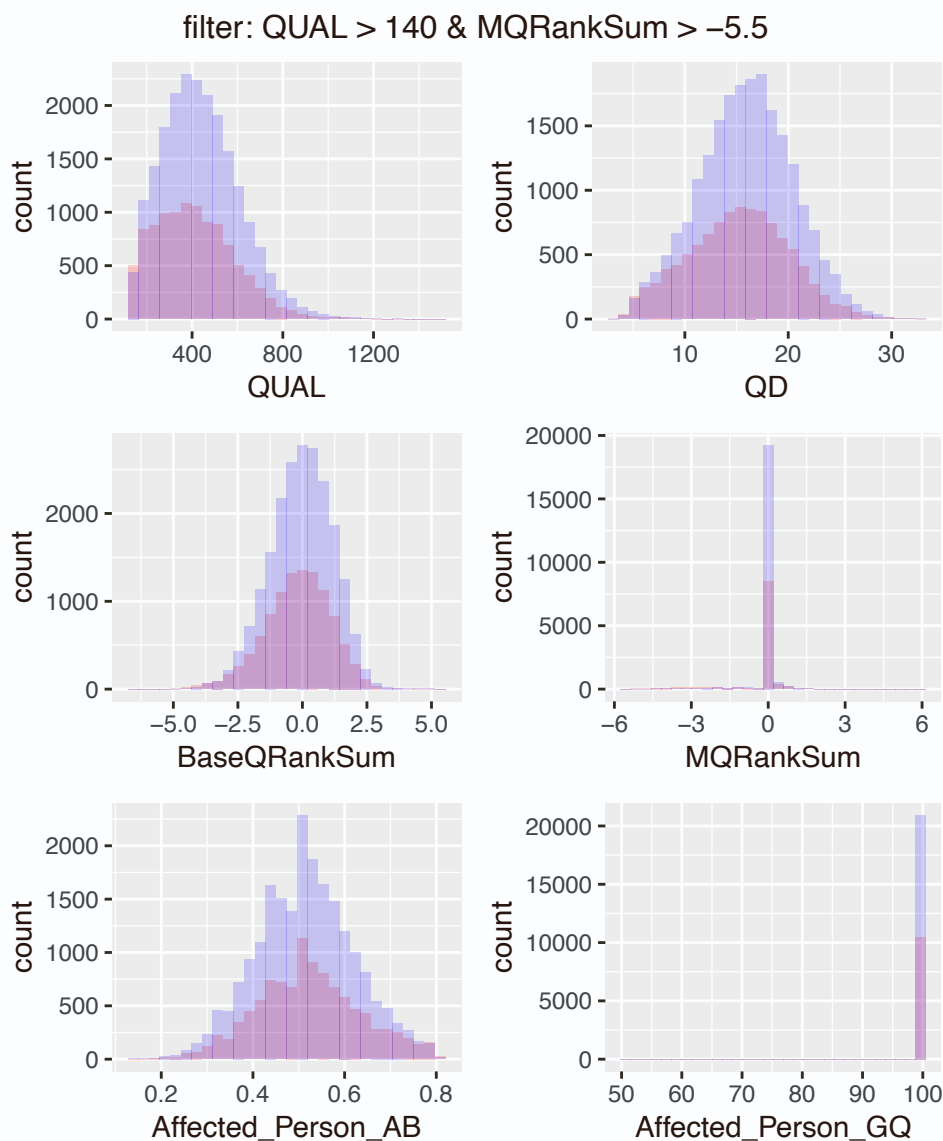


Figure S3. Histograms of quality scores of *de novo* mutations after filtering on QUAL and MQRankSum. Representation analogous to Figure S2: The individual histograms show quality scores of *variants de novo* mutations (post sample QC; intersect between Haplotype Caller and Unified Genotyper). However,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

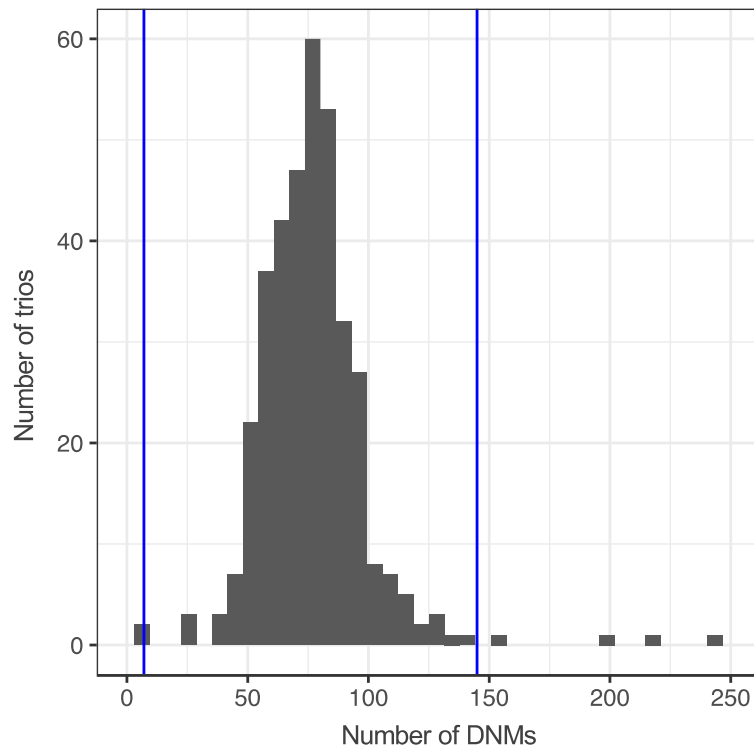


Figure S4. Number of *de novo* mutations per trio. The histogram shows the number of trios with the respective number of *de novo* mutations (DNMs; binwidth: 6.25). Trios with a number of DNMs above median + 3x IQR or below median - 3xIQR (blue line) were excluded for the following analyses. The cut-off 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.

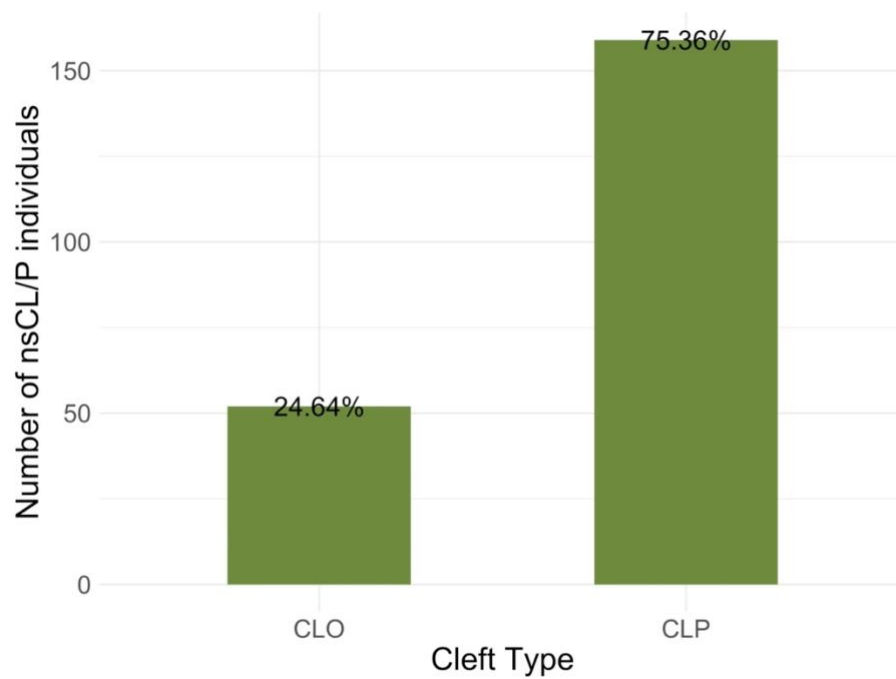


Figure S5. Distribution of different cleft phenotypes in 211 nsCL/P individuals.

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

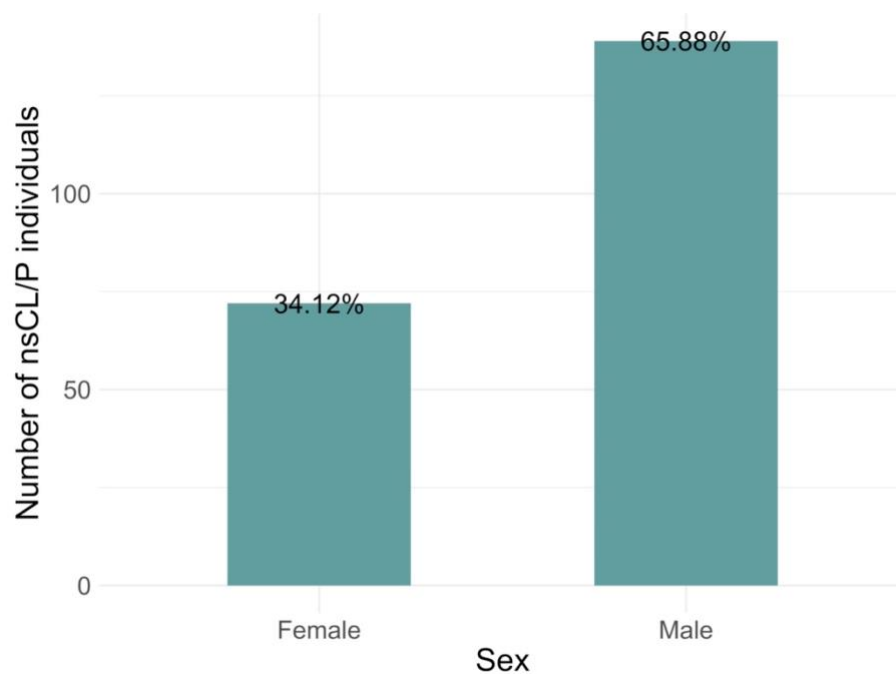


Figure S6. Distribution of sex in 211 nsCL/P individuals.

Abbreviation: nsCL/P – non-syndromic cleft lip with or without cleft palate

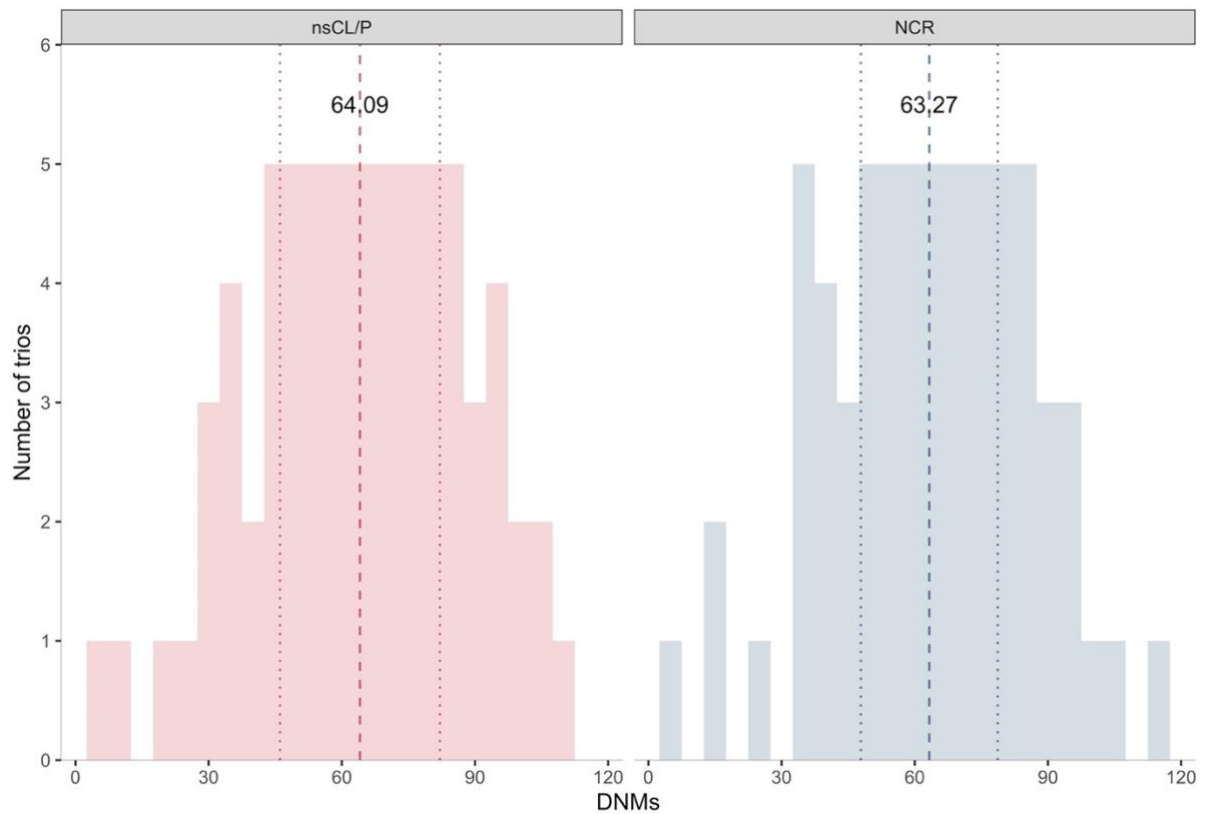


Figure S7. Number of *de novo* mutations per trio. Mean and standard deviation for number of *de novo* mutations (DNMs) are shown by lines (dashed: mean of DNMs per sample in cohorts, dotted: standard deviation of DNMs per sample in cohorts). Binwidth = 5, mean number of DNMs shown over dashed line.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

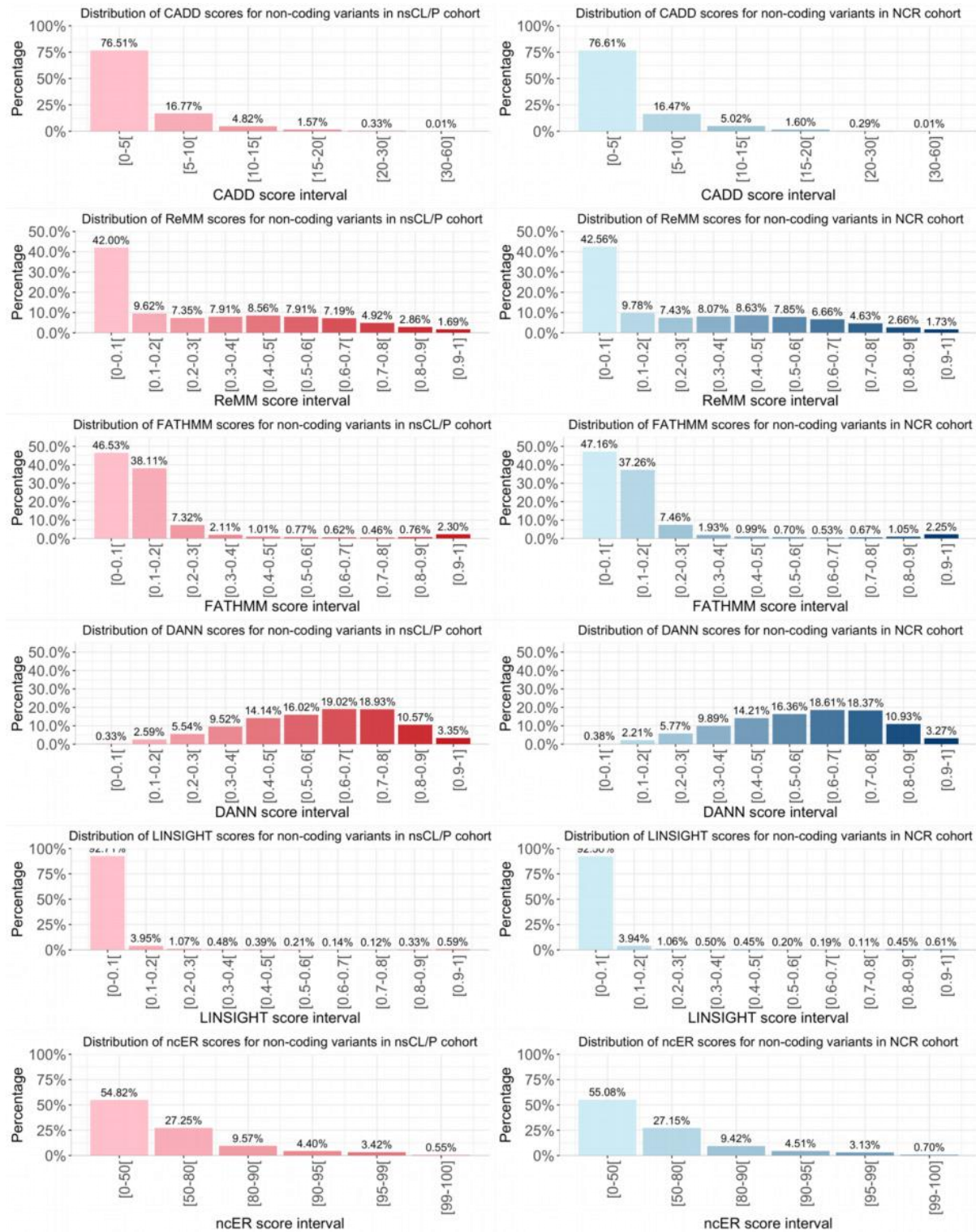


Figure S8. Distribution of prediction scores for non-coding *de novo* mutations in both cohorts. Distribution of six *in silico* prediction scores for non-coding *de novo* mutations (DNMs) in non-syndromic cleft lip with/without cleft palate (nsCL/P; red) and non-cleft reference cohort (NCR; blue). Thresholds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

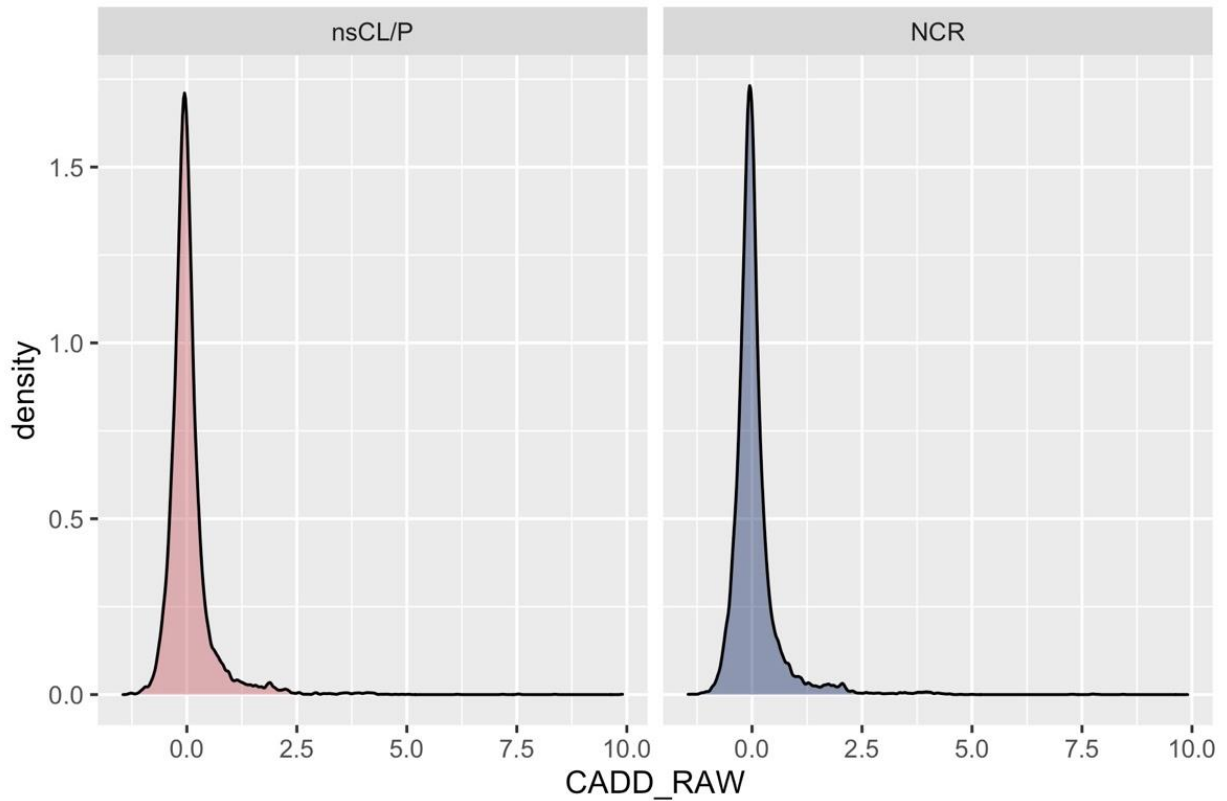


Figure S9. Cohort-wise distribution of raw CADD values for *de novo* mutations.

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

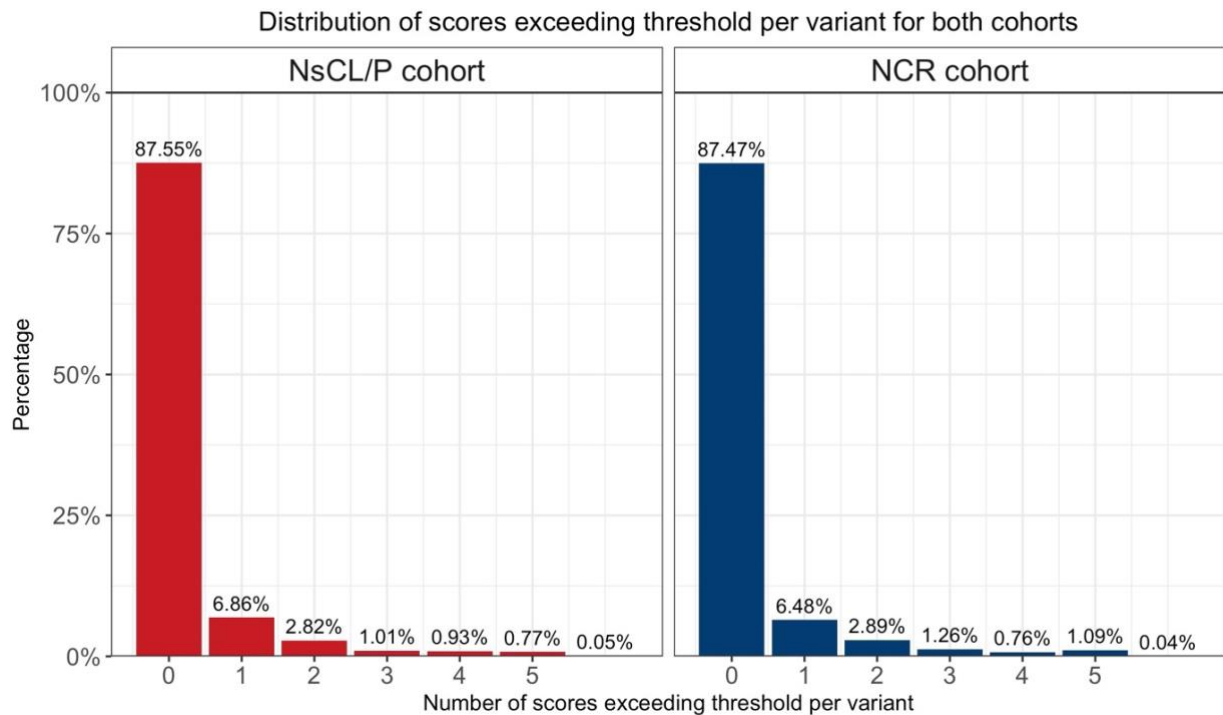


Figure S10. Percentage of *de novo* mutations exceeding the respective thresholds for the indicated number of *in silico* scores. Thresholds and references for six *in silico* prediction scores used for the comparison of cohorts are shown in Table S7.

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

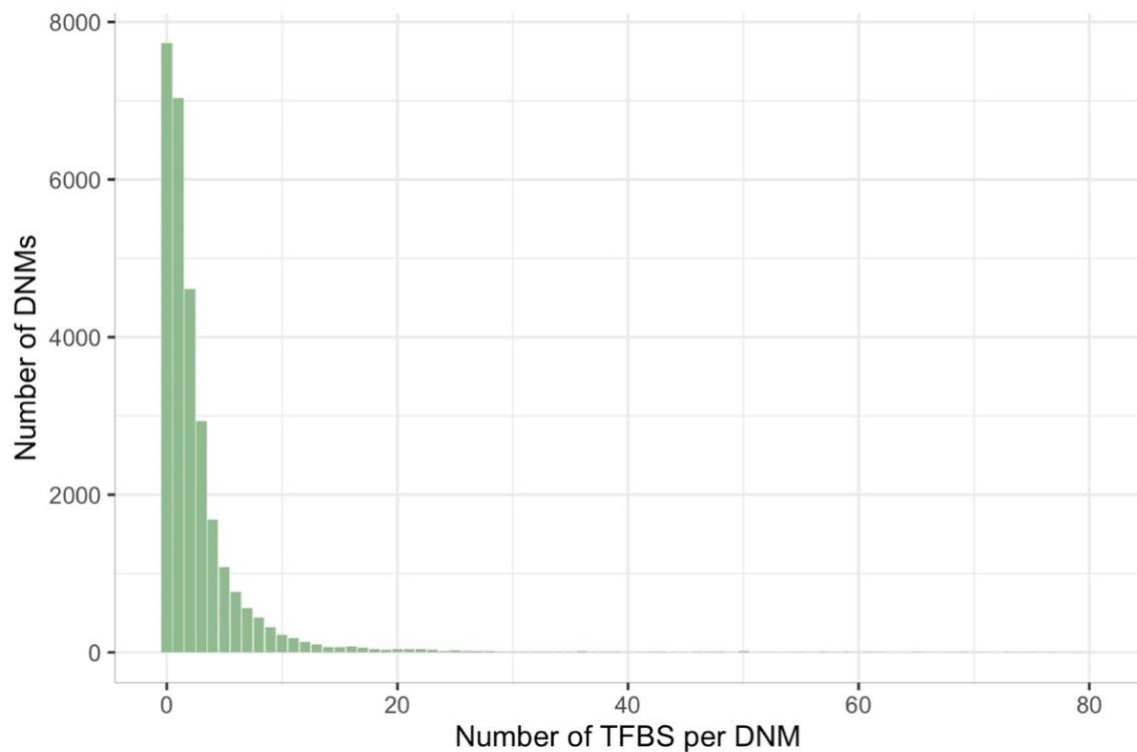


Figure S11. Number of predicted transcription factor binding sites per *de novo* mutation. For transcription factor binding sites (TFBS) identification, position weight matrix (PWM) information was compared to the genomic sequence around each DNM, with reference and alternative allele, using 810 PWMs 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

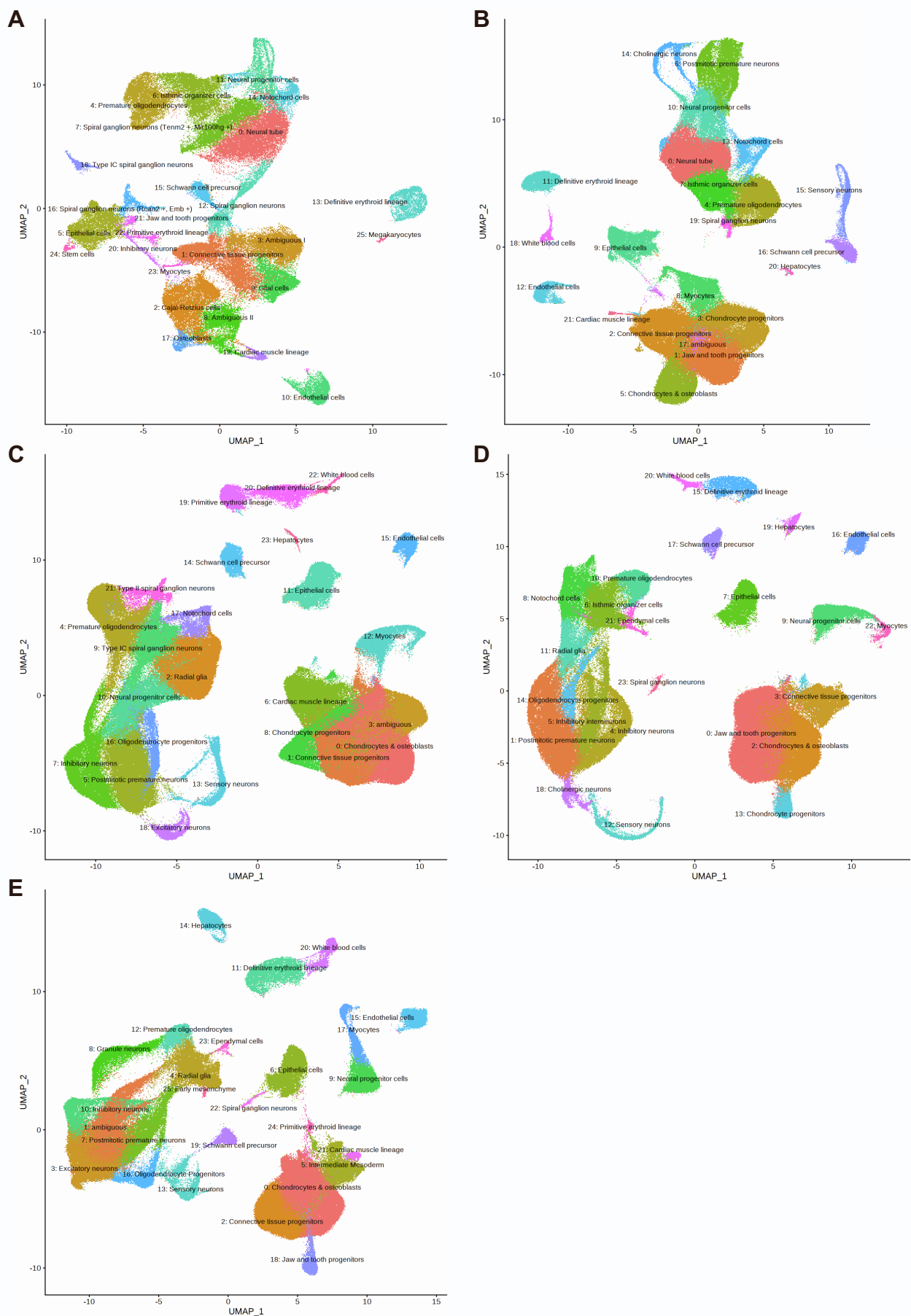


Figure S12. Single-cell data during murine embryogenesis. UMAP plots with cell clusters from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019) for embryonic days (A) E9.5, (B) E10.5, (C) E11.5, (D) E12.5, and (E) E13.5. The annotation of cell clusters is based on the original publication.

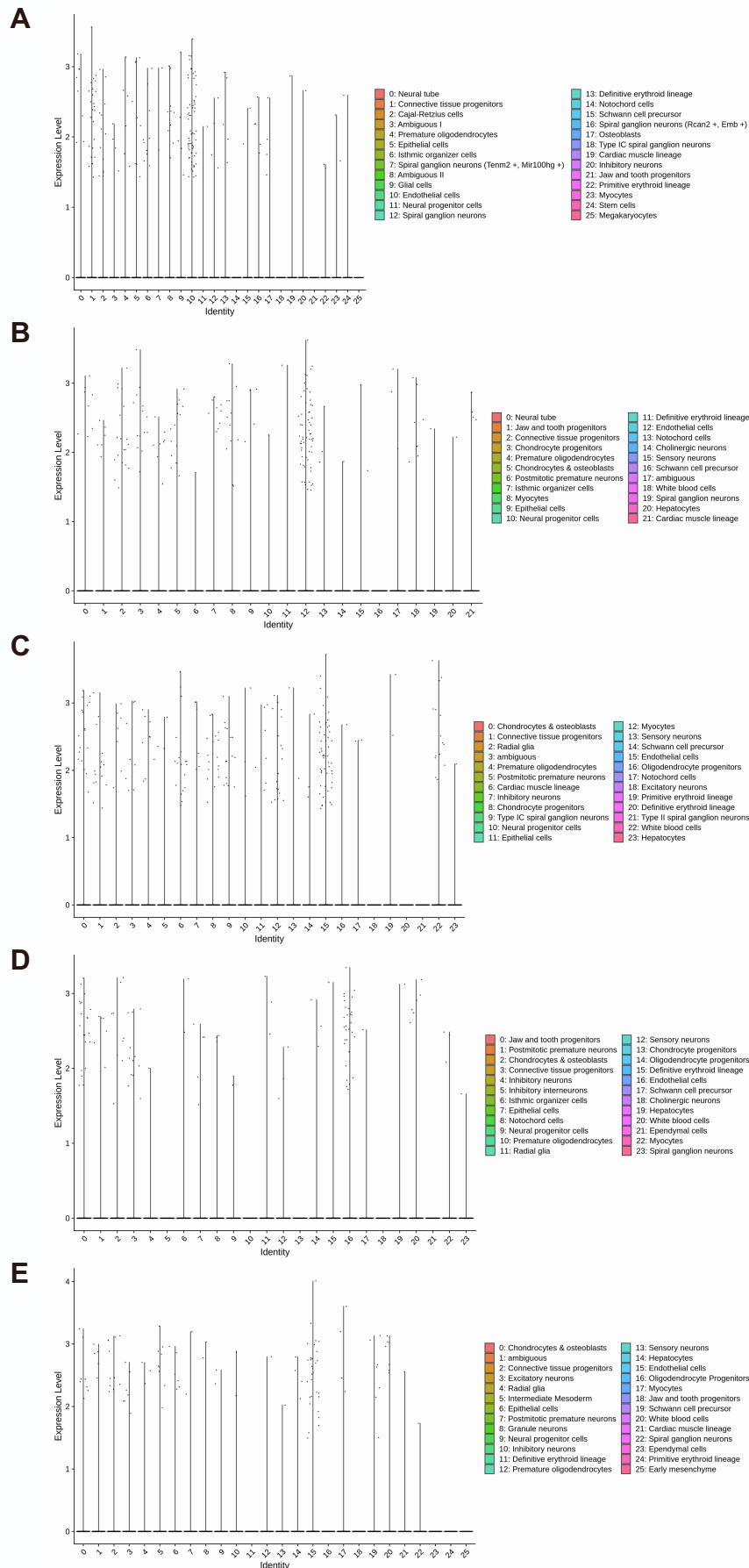


Figure S13. Expression of Activating Transcription Factor 3 in cell clusters from Mouse Organogenesis Cell Atlas at different embryonic days. Analysis of Activating Transcription Factor 3 (*Atf3*) expression on different days from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019): (A) E9.5, (B)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

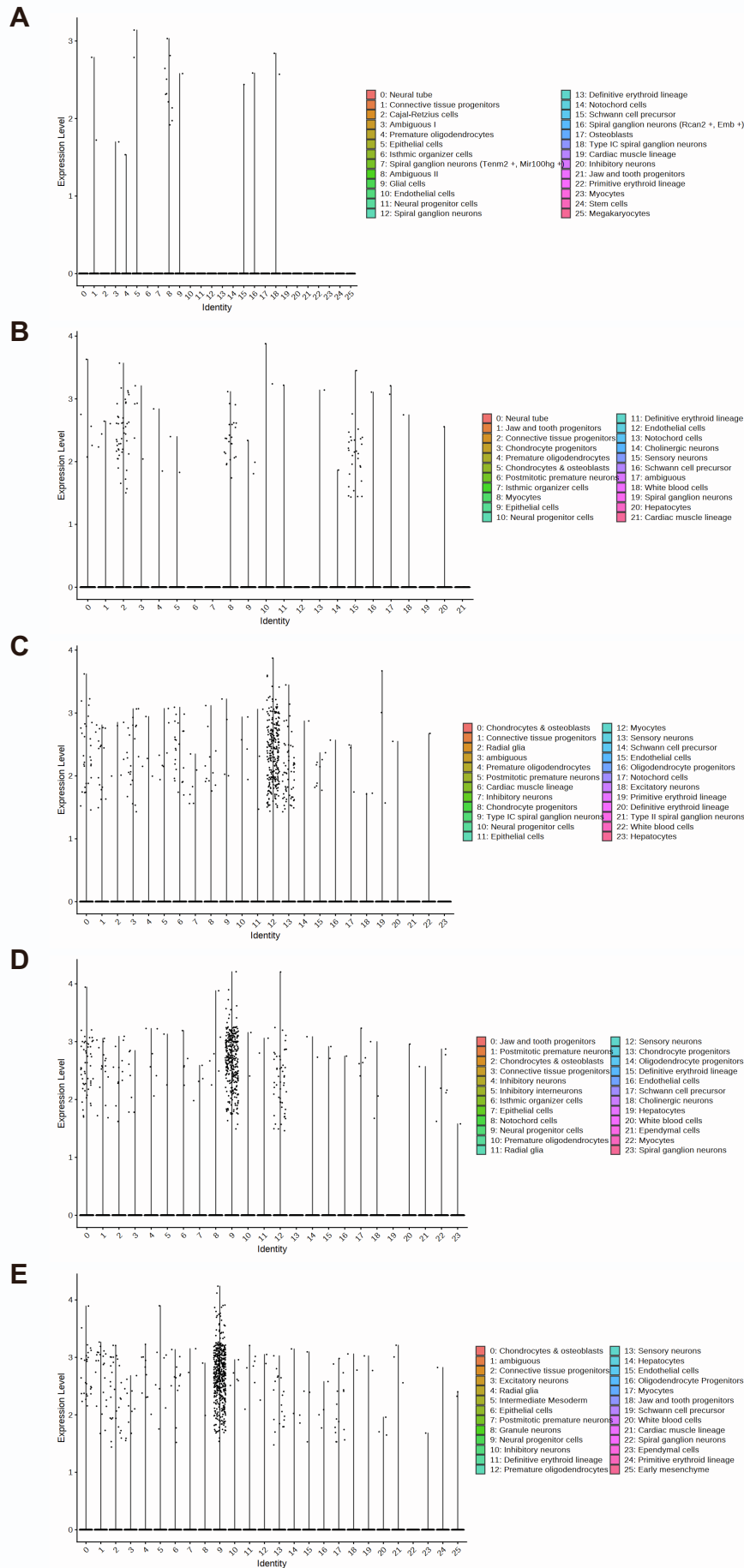


Figure S14. Expression of Musculin in cell clusters from Mouse Organogenesis Cell Atlas at different embryonic days. Analysis of Musculin (*Msc*) expression on different days from Mouse Organogenesis Cell Atlas (MOCA, Cao et al. 2019): (A) E9.5, (B) E10.5, (C) E11.5, (D) E12.5, and (E) E13.5. The annotation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

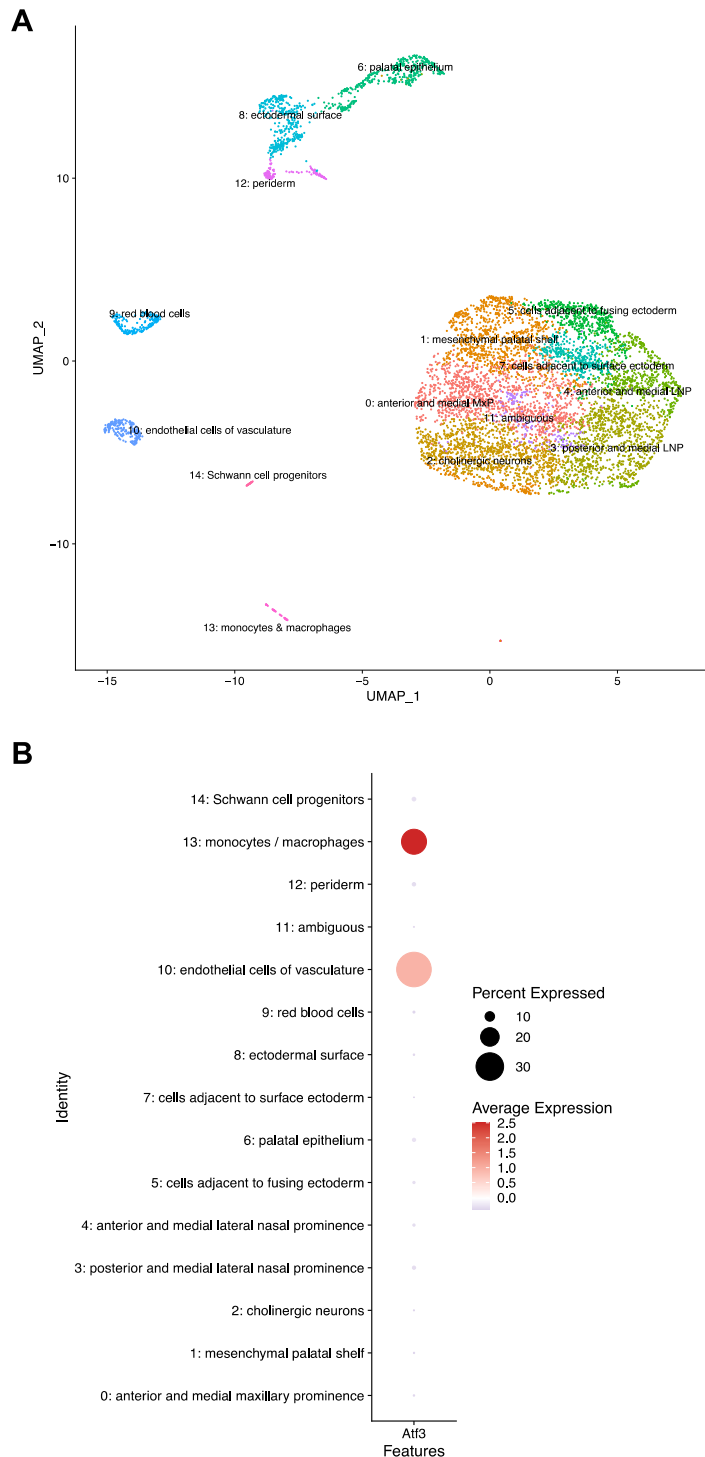


Figure S15. Single-cell expression data of the mouse lambdoidal junction at embryonic day E11.5. (A) Re-analysis of the single-cell data from Li et al. (2019) identified 15 cell clusters that are annotated based on marker gene expression. (B) Single-cell expression data of different cell clusters of 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

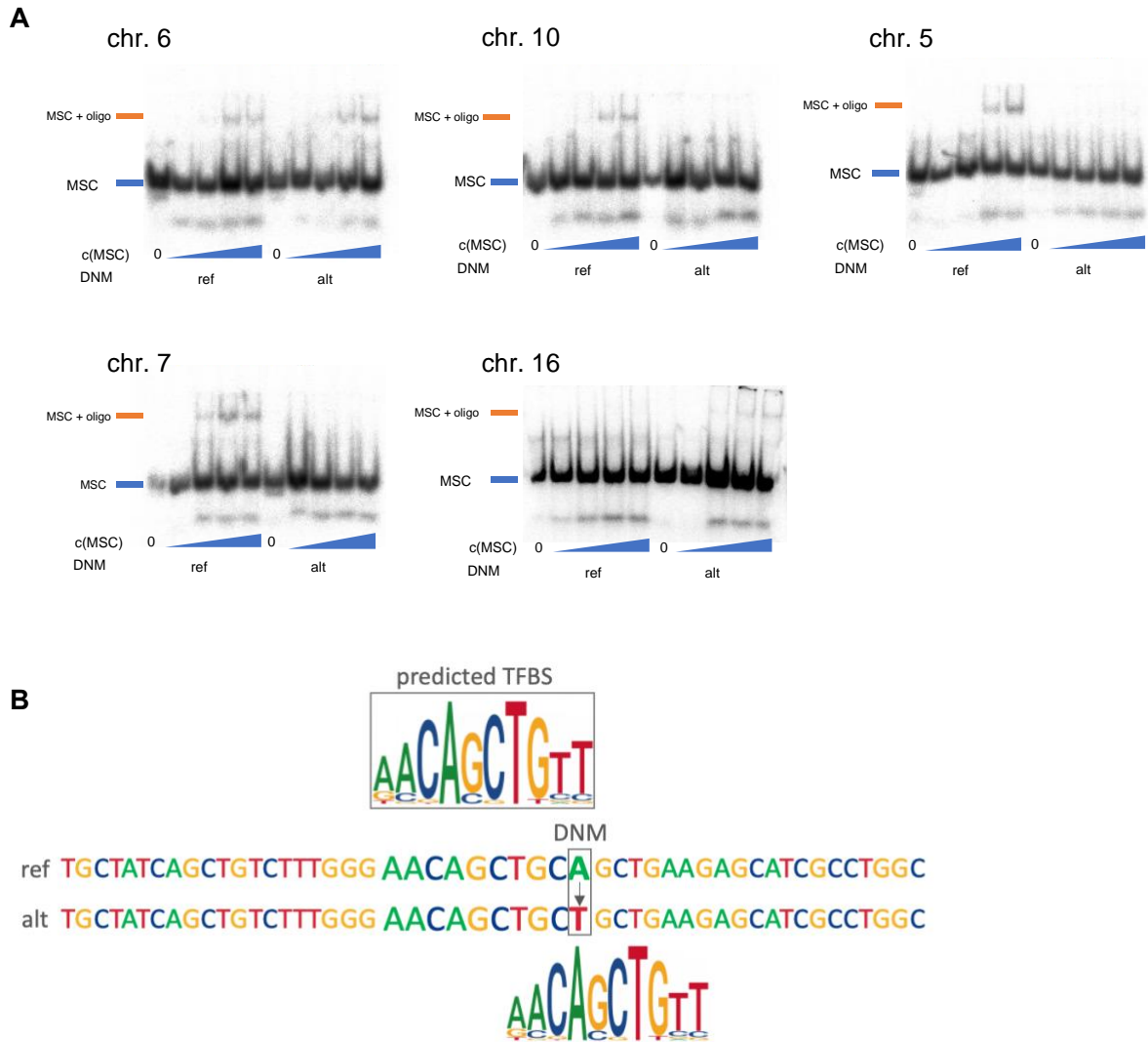


Figure S16. *In vitro* binding of Musculin using Electromobility Shift Assays.

(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

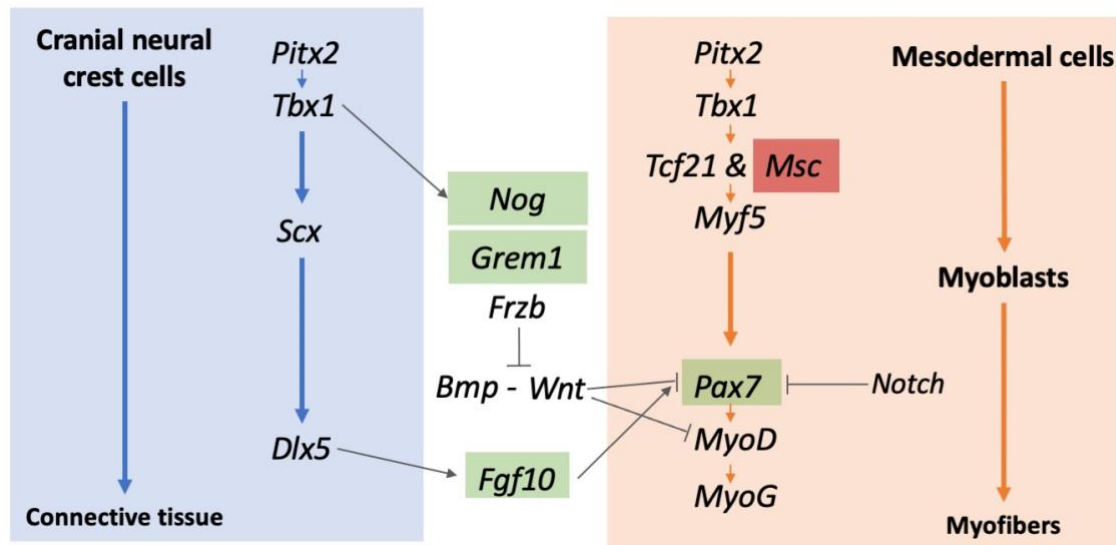


Figure S17. Schematic representation of genes involved in murine embryonic development of branchiomeric muscles. The interaction between cranial neural crest cells and mesodermal cells that develop into myofibers via myoblasts is illustrated by orange and blue backgrounds, respectively. Genes identified 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., 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.
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.