Untersuchung des Beitrags nicht-kodierender Neumutationen zur genetischen Ätiologie von nicht-syndromalen Lippen-Kiefer-Gaumenspalten

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Hohen Medizinischen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität

Bonn

Hanna Katharina Zieger

aus Aachen

2024

Angefertigt mit der Genehmigung

der Medizinischen Fakultät der Universität Bonn

1. Gutachterin: Prof. Dr. Kerstin U. Ludwig
2. Gutachter: Prof. Dr. Rayk Behrendt

Tag der Mündlichen Prüfung: 26.06.2024

Aus dem Institut für Humangenetik

Direktor: Prof. Dr. med. Markus M. Noethen

Inhaltsverzeichnis

Abkürzungsverzeichnis

Alt	Alternativ
CL/P	Lippen-Kiefer-Gaumenspalten (*cleft lip with/without cleft palate)*
cNCC	kraniale Neuralleistenzellen (*cranial neural crest cells)*
CNEs	konservierte, nicht-kodierende Elemente (*conserved non-coding elements*)
dbGaP	*database of Genotypes and Phenotypes*
DNM	Neumutation (*de novo* Mutation)
EMSA	*Electrophoretic Mobility Shift Assays*
GMKF	*Gabriella Miller Kids First*
GWAS	Genomweite Assoziationsstudien (*genome-wide association studies*)
hNCC	humane Neuralleistenzellen (*human neural crest cells*)
Indels	Insertionen/Deletionen
MOCA	*Mouse Organogenesis Cell Atlas*
MSC	Musculin
MWU	Mann-Whitney-U (statistischer Test)
nsCL/P	nicht-syndromale Lippen-Kiefer-Gaumenspalten (*non-syndromic cleft lip with/without cleft palate*)
PWM(s)	*position weight matrix(es)*
QF	Qualitätsfilter
Ref	Referenz
SNV(s)	Einzelnukleotidvariante(n) (*single nucleotide variant(s)*)
TADs	*Topologically associating domains*

TF(s)	Transkriptionsfaktor(en) (*transcription factor(s)*)
TFBS	Transkriptionsfaktorbindestellen
WES	Exomsequenzierung (*whole exome sequencing*)
WGS	Genomsequenzierung (*whole genome sequencing*)

1. Deutsche Zusammenfassung

1.1 Einleitung

1.1.1. Orofaziale Spalten

Orofaziale Spalten gehören zu den häufigsten angeborenen Fehlbildungen, mit einer Prävalenz von 1 : 700 Neugeborenen weltweit (Mossey und Modell, 2012). In Abhängigkeit vom Schweregrad der Erkrankung müssen sich die Betroffenen einer umfangreichen interdisziplinären Therapie unterziehen, welche meist vom ersten Lebensjahr bis in die Pubertät hineinreicht (Mossey et al., 2009; Zieger et al., 2023). Trotz der heute fortgeschrittenen Therapiemöglichkeiten besteht für Individuen mit orofazialen Spalten gegenüber der nicht-betroffenen Bevölkerung eine erhöhte Morbidität und Mortalität (Christensen et al., 2004). Die häufigste phänotypische Ausprägung der orofazialen Spalten ist die Form der Lippen-Kiefer-Gaumenspalten (*cleft lip with or without cleft palate*, CL/P). Dabei können entweder die Lippe allein (*cleft lip only*) oder Lippe und Gaumen (*cleft lip and palate*) betroffen sein. Neben der Klassifizierung anhand des phänotypischen Spektrums können orofaziale Spalten in „isolierte" (auch „nicht-syndromale") und „syndromale" Formen eingeteilt werden. Derzeit sind über 400 Krankheitsbilder bekannt, bei denen die orofaziale Spalte Bestandteil eines übergeordneten Syndroms ist (Jugessur et al., 2009). In Abgrenzung von diesen „syndromalen" Formen wird das Auftreten von orofazialen Spalten ohne weitere Symptome als „isolierte" oder „nicht-syndromale" Form bezeichnet. Den größten Teil der orofazialen Spalten bilden Individuen mit nicht-syndromalen CL/P (nsCL/P), mit einer weltweiten Inzidenz von ca. 1 : 1.000 (Mangold et al., 2011).

1.1.2. Ätiologie von nsCL/P

NsCL/P gehören zu den multifaktoriellen Erkrankungen, deren Auftreten genetische Faktoren und Umweltfaktoren zugrunde liegen. Als relevante Umweltfaktoren für die Entstehung von nsCL/P konnten unter anderem maternales Rauchen (Little et al., 2004) und ein Mangel von Vitamin B-Komplex und Vitamin A in der Schwangerschaft (Jugessur und Murray, 2005) identifiziert werden, wobei unabhängige Replikationen sowie der Nachweis von kausalen Zusammenhängen noch ausstehen.

Die Gesamt-Heritabilität, welche bei multifaktoriellen Erkrankungen durch epidemiologische Studien wie z. B. Zwillingsstudien geschätzt werden kann (Visscher et al., 2008), beträgt bei nsCL/P etwa 90 % (Grosen et al., 2011). Dabei ist der genetische Beitrag komplex, d. h. es tragen sowohl häufige als auch seltene genetische Varianten zur Ätiologie bei (Thieme und Ludwig, 2017).

Häufige Risikovarianten (*common variants*) treten auch in vielen gesunden Individuen auf und haben demnach eine geringe Penetranz – ihr Auftreten bedingt nicht zwingend das Auftreten der Erkrankung (Ludwig et al., 2019). Die Untersuchung häufiger Varianten kann systematisch durch genomweite Assoziationsstudien (*genome-wide association studies*, GWAS) erfolgen. Bei GWAS werden Einzelbasenaustausche, sogenannte *single nucleotide variants* (SNVs), unter Einsatz von kostengünstigen SNV-Arrays in Erkrankten sowie nicht-betroffenen Individuen genotypisiert, d. h. die allelische Ausprägung ermittelt. Durch einen Vergleich der Allelfrequenzen zwischen beiden Gruppen können SNVs in bestimmten Regionen („Loci") mit der untersuchten Erkrankung assoziiert werden (Uffelmann et al., 2021).

Bis heute konnten durch GWAS ca. 45 Risikoloci für nsCL/P identifiziert werden (Beaty et al., 2010; Birnbaum et al., 2009; Leslie et al., 2016; Ludwig et al., 2012; Ludwig et al., 2017; Mangold et al., 2010; Welzenbach et al., 2021; Yu et al., 2017). Diese sind fast ausschließlich in nicht-kodierenden Abschnitten des Genoms lokalisiert (Thieme und Ludwig, 2017). Insgesamt wird der Beitrag häufiger Varianten an der Heritabilität von nsCL/P in europäischen Individuen auf ca. 30 % geschätzt (Welzenbach et al., 2021). Neben diesen häufigen SNVs tragen auch seltene oder niedrigfrequente Varianten zur Ätiologie von nsCL/P bei (Thieme und Ludwig, 2017). Seltene Varianten können zwar über dieselben biologischen, nicht-kodierenden Mechanismen wirken, haben aber möglicherweise größere Effektstärken und eignen sich daher besser für funktionelle Folgeuntersuchungen (Fakhouri et al., 2014; Rainger et al., 2014). Neben SNVs gehören auch strukturelle Varianten (bspw. Kopienzahlvarianten, Insertionen/Deletionen (sogenannte „Indels") oder Translokationen) zum genetischen Spektrum der nsCL/P (Basha et al., 2018; Lansdon et al., 2023).

Für die Identifizierung von seltenen SNVs und strukturellen Varianten eignen sich die Technologien des *next generation sequencing*: Exomsequenzierung (*whole exome*

sequencing, WES) oder Genomsequenzierung (*whole genome sequencing,* WGS). WGS-Analysen, die neben den im Rahmen von WES sequenzierten Genen den ganzen nicht-kodierenden Bereich abdecken und dadurch eine deutlich größere Anzahl von detektierten Varianten pro individuelles Genom generieren, sind u. a. aufgrund der Größe der generierten Datenmenge und den mit Speicherung, Verarbeitung sowie Auswertung verbundenen Kosten (Ludwig et al., 2019), aber auch aufgrund der ethischen Problematik solcher sensiblen Datensätze (McGuire et al., 2008) auch heutzutage für viele Erkrankungen nur für kleine klinische Kohorten verfügbar.

Der Anteil häufiger und seltener Varianten an der genetischen Ätiologie ist abhängig von der untersuchten Erkrankung: Bei Erkrankungen, die mit einer erniedrigten Reproduktionsrate einhergehen, ist aufgrund des Selektionsdrucks zu erwarten, dass die Frequenz kausaler Mutationen durch evolutionäre Prozesse niedrig gehalten wird. Dadurch kann von einem größeren Anteil neuerer, seltener Varianten mit im Schnitt größeren Effektstärken ausgegangen werden (Raychaudhuri, 2011). Epidemiologische Daten zur Sterblichkeit von Patienten mit orofazialen Spalten in Ländern ohne westliche Gesundheitsversorgung weisen auf einen solchen Selektionsdruck hin. In Ländern ohne Verfügbarkeit einer operativen Therapie zeigt sich eine hohe Sterblichkeitsrate bei jungen Betroffenen (Abbildung 1); (Christensen et al., 2004; Mossey und Modell, 2012). Durch WES bzw. gezielte Sequenzierung von GWAS Kandidaten-Genen konnten einige Studien bereits einen Beitrag seltener Varianten zur Ätiologie von nsCL/P belegen (Basha et al., 2018; Butali et al., 2014; Letra et al., 2014; Savastano et al., 2017; Sylvester et al., 2020).

In der Ätiologie von nsCL/P könnten auch *de novo* Mutationen (DNMs), eine spezielle Form seltener Varianten, die neu in betroffenen Individuen auftreten, eine Rolle spielen. Diese Hypothese wird unterstützt durch weitestgehend stabile Prävalenzen von nsCL/P, trotz der erhöhten Mortalität von betroffenen Individuen mit fehlender operativer Versorgung (Abbildung 1). Zudem zeigt sich ein erhöhtes Erkrankungsrisiko für die Nachfahren eines Betroffenen, im Vergleich zu vorhergehenden Generationen (Grosen et al., 2010). Für die Detektion solcher DNMs sind sowohl die genetischen Daten des betroffenen Individuums als auch beider Elternteile notwendig (sogenannte Trios).

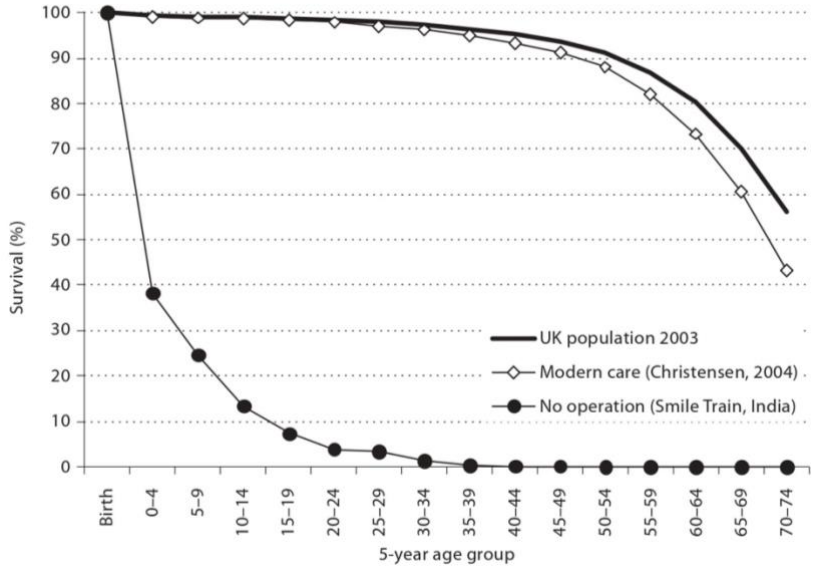

Abb. 1: Langzeit-Überlebenskurve für Individuen mit orofazialen Spalten mit und ohne operative Therapie im Vergleich zur Normalbevölkerung.

Dargestellt sind drei Überlebenskurven für (i) die Population in Großbritannien (UK) im Jahr 2003 (graue Linie), für (ii) Individuen mit orofazialen Spalten mit Zugang zu westlicher Gesundheitsversorgung (Linie mit weißen Rauten-förmigen Punkten) und für (iii) Individuen mit orofazialen Spalten ohne Zugang zu operativer Versorgung (Linie mit schwarzen Punkten). Dabei ist auf der y-Achse der Anteil der Überlebenden in Prozent gegen das Alter in 5-Jahres-Intervallen auf der x-Achse dargestellt. Hier zeigt sich eine deutlich erhöhte frühkindliche Mortalität bei fehlendem Zugang zu operativer Versorgung, jedoch auch eine erhöhte Mortalität nach Spaltkorrektur im Lebensverlauf. Diese Abbildung ist mit Genehmigung von S. Karger AG, Basel, aus Mossey und Modell (2012) entnommen.

1.1.3 Das nicht-kodierende Genom

Trotz der Vielzahl identifizierter Risikoloci für nsCL/P, besonders aus dem Frequenzspektrum der häufigen Varianten, ist das Verständnis der zugrunde liegenden Pathophysiologie bisher nur eingeschränkt möglich. Dies liegt einerseits in der unzureichend geklärten Funktion der nicht-kodierenden Bereiche, sowie andererseits in der geringen Effektstärke der Risikovarianten begründet. Man geht davon aus, dass nicht-kodierende Risikovarianten in funktionell aktiven regulatorischen Elementen liegen, welche zu einer zeit- und gewebespezifischen Expression von benachbarten Genen beitragen (Buecker und Wysocka, 2012; Bulger und Groudine, 2010). Um innerhalb großer Datensätze ätiologisch relevante nicht-kodierende Varianten zu identifizieren, können experimentelle Designs bzw. die Integration bereits verfügbarer

funktioneller Datensätze genutzt werden, um Varianten mit regulatorischen Effekten herauszuarbeiten, z. B. zelltypspezifische Chromatin-Modifikationen oder *chromatin conformation capture assays* für die Darstellung der räumlichen 3D-Anordnung des Chromatins. In Form von Hi-C Assays kann größer eine Einteilung des gesamten Genoms in sogenannte *topologically associating domains* (TADs) erfolgen, die gewebespezifische funktionale Einheiten des Genoms abbilden sollen (Dixon et al., 2012; Lieberman-Aiden et al., 2009).

Eine Herausforderung bei der Untersuchung der Ätiologie von nsCL/P mit Hilfe solcher Daten besteht darin, dass die Entwicklung der Gesichtsstrukturen während der Embryonalentwicklung stattfindet und humanes Gewebe aus diesem Stadium nur begrenzt zur Verfügung steht (Thieme und Ludwig, 2017). Da nur einzelne Datensätze von humanem kraniofazialen Gewebe (bspw. Wilderman et al., 2018) verfügbar sind, wird für molekulare Untersuchungen auf geeignete *in vitro* Zellsysteme wie humane Neuralleistenzellen (*human neural crest cells*, hNCC; Rada-Iglesias et al., 2012) und kraniale Neuralleistenzellen von Menschen und Schimpansen (*cranial neural crest cells*, cNCC; Prescott et al., 2015) zurückgegriffen.

1.1.4 Zielsetzung der Arbeit

Der Beitrag seltener Varianten in nicht-kodierenden Regionen zu nsCL/P wurde bisher nicht systematisch untersucht. Das Ziel dieser Doktorarbeit war es daher, mit Hilfe eines WGS-Trio-Datensatzes mögliche kausale DNMs in nicht-kodierenden Bereichen des Genoms zu identifizieren und diese durch Integration weiterer Datenebenen zu priorisieren.

1.2 Material und Methoden

1.2.1 Datengrundlage

Grundlage der Arbeit waren zwei WGS-Datensätze des *Gabriella Miller Kids First* (GMKF) Programms (im Folgenden GMKF Projekt). Nach Qualitätskontrolle (siehe Tabelle 1) umfassten diese Datensätze zum einen Individuen mit nsCL/P und ihre Eltern (n = 220), zum anderen Individuen mit Ewing Sarkom und deren Eltern (n = 289, Kontrollkohorte).

Durch dieses „Trio“-Design wurde die genomweite Untersuchung von DNMs ermöglicht und erfolgte ausschließlich an Individuen europäischer Herkunft.

Tab. 1: Übersicht der verwendeten Genomsequenz-Datensätze

Dargestellt sind die beiden Datensätze mit Projekttitel, die antizipierte Anzahl an Individuen des Projekts, die vollständige Anzahl an verwendeten Trios (Datensätze von WGS-Daten für Individuum und beide Elternteile) sowie die Anzahl an Trios nach Qualitätsfiltern (QF) für DNM-Anzahl pro Trio.

Abkürzungen: GMKF - *Gabriella Miller Kids First*; nsCL/P - nicht-syndromale Lippenspalte mit/ohne Gaumenspalte; QF - Qualitätsfiltern; WGS - Genomsequenzierung

	nsCL/P Kohorte	**Kontroll Kohorte**
GMKF Projekttitel (2015 X01)	“Genomic Studies of Orofacial Clefts Birth Defects”	Genetic Contribution to Ewing Sarcoma in 330 parent-Offspring Trios”
im Datensatz vorhandene Samplezahl	1.242	1.112
vollständige Trios	332	289
Filter Phänotyp, gesunde Eltern:	220	
finale Anzahl Trios (nach QF)	211	284

Die WGS-Daten wurden nach einem genehmigten Datenzugangs-Antrag über die *NIH database of Genotypes and Phenotypes* (dbGaP; https://www.ncbi.nlm.nih.gov/gap, Tryka et al., 2014) abgerufen. Das *Alignment* der Rohdaten an das Referenzgenom GRCh37 sowie das anschließende *Variant Calling*, bei dem nur autosomale DNMs ermittelt wurden, erfolgte in Zusammenarbeit mit Manuel Holtgrewe und Dieter Beule vom *Berlin Institute of Health* bzw. dem *Max Delbrück Center* in Berlin. Um eine hohe Qualität des Datensatzes zu gewährleisten, wurden nur Trios mit einer Anzahl DNMs ≤ Median ± 3. Interquartilsabstand eingeschlossen und abschließend die DNMs anhand von zwei Qualitäts-Scores gefiltert. Für Details zu diesen Arbeitsschritten sei auf die Publikation und deren *Supplemental Methods* (S. 24) verwiesen.

1.2.2 Annotation und deskriptive Analysen

Für alle DNMs erfolgte zunächst eine Klassifikation anhand ihrer genomischen Lokalisation als exonisch, intronisch bzw. intergenisch, sowie eine Annotation mit der Allelfrequenz aus gnomAD v3.1 (Karczewski et al., 2020) und sechs verschiedenen *in silico* Prädiktions-Scores (z. B. CADD (Kircher et al., 2014), siehe Tabelle S7 aus *Supplement* der Publikation). Diese *in silico* Prädiktions-Scores nutzen verschiedene Ebenen an Informationen (z. B. zu Konservierung, Genregulation oder Populationsfrequenz) um vorherzusagen, wie pathogen bzw. schädigend (oder nicht) einzelne genetische Varianten sind. Anschließend wurde getestet, ob signifikante Unterschiede in der Anzahl an DNMs pro Trio, deren prozentualer Verteilung in den genomischen Abschnitten oder bezüglich der *in silico* Prädiktions-Scores bestehen.

1.2.3 DNM-Enrichment-Analysen

Im nächsten Schritt wurden DNMs aufgrund gemeinsamer genomischer Eigenschaften zusammengefasst. Dabei war die Hypothese, ätiologisch relevante DNMs nicht durch genomische Nähe, sondern aufgrund ihrer Anreicherung (sog. „Enrichment“) in Regionen mit gewebespezifischen funktionellen Effekten zu detektieren (Abbildung 2). Als statistische Methode kam das FunciVar package (Jones et al., 2020) zum Einsatz (siehe *Supplemental Methods* S. 26).

Für die Enrichment-Analysen wurden zwei Ansätze durchgeführt. Für den ersten Ansatz wurden Regionen mit *a priori* Evidenz für eine Rolle in der kraniofazialen Entwicklung ausgewählt: bspw. Enhancer-Elemente aus dem VISTA Enhancer Browser (Visel et al., 2007) mit gewebespezifischer Aktivität und ein Datensatz mit konservierten genomischen Regionen (*conserved non-coding elements*, CNEs; Short et al., 2018). Weiter wurden die DNMs nach epigenetischen Zuständen der genomischen Region auf Basis von *Chromatin States* aus drei Gewebetypen, die relevant für die Gesichtsentwicklung sein könnten, eingeteilt: cNCC (Prescott et al., 2015), hNCC (Rada-Iglesias et al., 2012), und humanes embryonales kraniofaziales Gewebe (Wilderman et al., 2018). Als zweiter Ansatz erfolgte eine Einteilung der DNMs in TADs (basierend auf embryonalen Stammzellen, n = 2.991; Dixon et al., 2012). Zusätzlich wurde eine Subgruppe von TADs erstellt, in denen sich

häufige Risikovarianten für nsCL/P aus den bisherigen GWAS-Analysen befanden ($TADs_{GWAS}$).

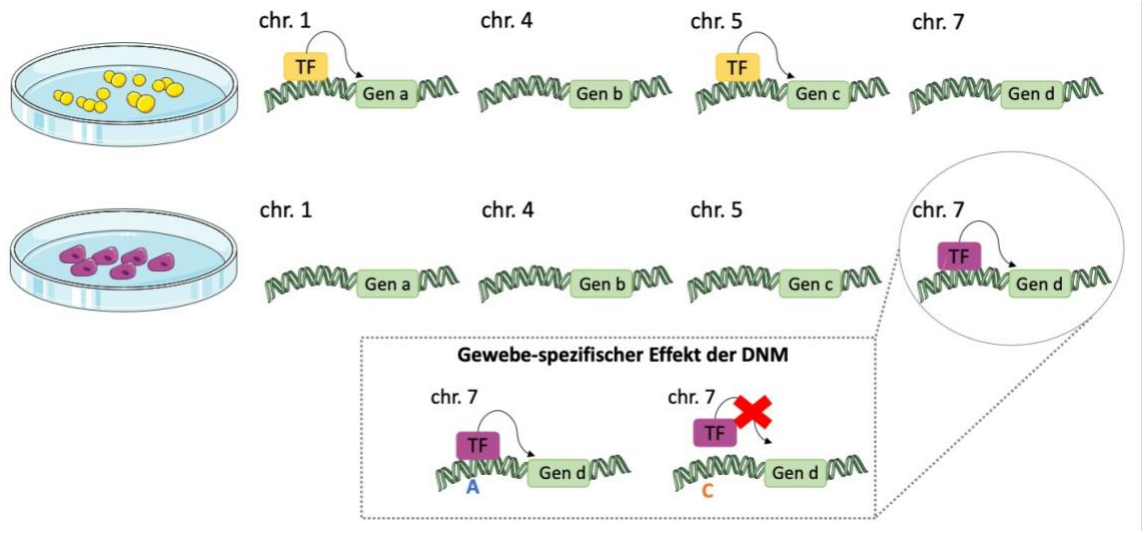

Abb. 2: Identifizierung von ätiologisch relevanten DNMs am Beispiel der zelltyp-abhängigen regulatorischen Funktionen einer genomischen Region.

Ein Transkriptionsfaktor (TF) bindet an bestimmte genomische Regionen (dargestellt durch Abschnitte auf Chromosom (chr.) 1, 4, 5, und 7). Durch diese Bindung wird durch den Transkriptionsfaktor anschließend die Expression benachbarter Gene aktiviert. Diese Bindung und darauffolgende Aktivierung der Genexpression finden nur in bestimmten Geweben statt. Hier sind die genomischen Regionen für zwei Gewebetypen (Petrischalen mit verschiedenfarbigen Zellen) in zwei Reihen dargestellt. Tritt eine *de novo* Mutation (DNM) innerhalb eines solchen Bindungsbereich auf, kann der Transkriptionsfaktor eventuell nicht binden und somit die gewebespezifische Aktivität verändern. Teile der Abbildung wurden mit Hilfe von Bildern von Servier Medical Art erstellt. Servier Medical Art von Servier ist unter der Creative Commons Attribution 3.0 Unported License lizenziert (https://creativecommons.org/licenses/by/3.0/)."

1.2.4 Untersuchung von allelischen Effekten auf Transkriptionsfaktorbindestellen

Eine mögliche veränderte Bindung von Transkriptionsfaktoren (*transcription factors*, TFs) wurde durch die Auswertung von Transkriptionsfaktorbindestellen (TFBS) an DNM-Positionen mittels des Tools denovoLOBGOB (Short et al., 2018) untersucht. Dabei werden die Effekte von DNMs auf Transkriptionsfaktor- (TF) Bindungen basierend auf der genomischen Sequenz um die DNM-Position und einer *position weight matrix* (PWM) als Bindeprofil berechnet (siehe Abbildung 3).

Insgesamt wurden für alle SNVs im DNM-Datensatz mögliche Veränderungen der Bindungswerte berechnet, dabei wurden 810 verschiedene PWMs aus der JASPAR 2020 Datenbank betrachtet (Fornes et al., 2020). Die Differenz der jeweiligen Bindungswerte zwischen Referenz- (Ref) und Alternativ- (Alt) Allel gibt die Bindungsänderung (*binding change*, BC) des Transkriptionsfaktors durch die DNM an. Hierbei können Bindungsstärkenzunahme (*gain of binding*), Bindungsstärkenreduktion (*loss of binding*) und unveränderte Bindungsstärke (*silent effect*) durch die DNM unterschieden werden. Für eine Beschreibung der Änderungen in denovoLOBGOB und des Filterns der ausgegebenen TFBS sei auf die *Supplemental Methods* (S. 28 f.) verwiesen.

Anschließend wurde jedes Bindemotiv auf einen möglichen Unterschied (i) in der Anzahl der absoluten Bindungsereignisse (*Hits,* Fisher`s Exact Test), und (ii) der quantitativen Veränderung der Bindungsstärke (Mann-Whitney-U (MWU) Test) getestet. Die jeweiligen Effektstärken wurden mittels eines *Fold Change* bestimmt und Kandidaten-TFs ausgewählt (siehe *Supplemental Methods* S. 29).

Position weight matrix (PWM)

	1	2	3	4	5	6
A	-2.27	-4.41	-4.41	-4.24	-0.50	**0.93**
C	0.08	-4.68	-4.68	-4.29	**1.63**	-0.82
G	-3.29	-2.53	**1.96**	**1.94**	-3.43	-0.54
T	**1.40**	**1.90**	-5.08	-4.26	-3.41	-0.26

Abb. 3: Beispiel eines Transkriptionfaktorbindeprofils

Dargestellt sind das Matrix Profil MA0161.1 für den Transkriptionsfaktor (TF) „Nuclear Factor 1 C-Type" (oben) mit zugehöriger PWM (unten). Diese sind der JASPAR 2020 Datenbank entnommen. Die Matrix bildet die Konsensussequenz ab, an die der TF präferentiell bindet. Für jede der sechs Positionen des Bindeprofils enthält die PWM einen Wert für die vier möglichen Basen. Mit dieser PWM kann für jede genomische Sequenz mit einer Länge von sechs Basen ein Wert für die Bindungsstärke des

Transkriptionsfaktors berechnet werden. Die Abbildung wurde unter Verwendung der Position Frequency Matrix aus der JASPAR 2020 (Fornes et al., 2020) open-access Datenbank erstellt, die unter der Creative Commons Attribution 4.0 International License lizenziert ist. (https://creativecommons.org/licenses/by/4.0/)

Abkürzungen: PWM - *position weight matrix*; TF - Transkriptionsfaktor

1.2.5 Durch Kollaborationspartner zur Verfügung gestellte Daten

a) *Single-cell* Expressionsdatensätze für identifizierte Kandidaten-TFs

Zur Priorisierung von Kandidaten-TFs für eine experimentelle Aufarbeitung wurde in Zusammenarbeit mit Anna Siewert (AG Ludwig, Bonn) die Expression der Gene dieser TFs in Gewebetypen, die zeitlich und räumlich für die Entwicklung von nsCL/P relevant sind, analysiert (Siewert et al., 2023). Dazu wurden zwei *single-cell* Expressionsdatensätze verwendet: der *Mouse Organogenesis Cell Atlas* (MOCA) (Mausembryonen von fünf verschiedenen Embryonaltagen (E9.5, E10.5, E11.5, E12.5, und E13.5); Cao et al., 2019) sowie ein Datensatz der lambdoidalen Verbindung (*lambdoidal junction*) im Mausgesicht am Embryonaltag 11.5 (Li et al., 2019). Für genauere Informationen zur Re-Analyse der beiden Datensätze sei auf die *Supplemental Methods* (S. 29 f.) verwiesen.

b) *Electrophoretic Mobility Shift Assays*

Zur Validierung der vorhergesagten TFBS an DNM-Positionen wurde in Zusammenarbeit mit Katrin Paeschke und Stefan Juranek (Uniklinikum Bonn) die DNA-Protein-Interaktion des Transkriptionsfaktors Musculin (MSC) mit der jeweiligen genetischen Sequenz mit Hilfe von *Electrophoretic Mobility Shift Assays* (EMSA) untersucht. Hierzu wurde das Bindungsmotiv an der DNM-Position ± 20 Basenpaare mit Ref- und Alt-Allel mit fünf verschiedenen Konzentrationen von MSC inkubiert. Für jede getestete DNM-Bindungsreaktion wurden je drei Wiederholungen für Ref- und Alt-Allel durchgeführt (siehe *Supplemental Methods* S. 30). In dieser deutschen Zusammenfassung wird das Vorgehen für eine Transkriptionsfaktorbindestelle gezeigt, die gesamten Ergebnisse der EMSA-Versuche sind in Tabelle S30 des *Supplements* der Publikation zu finden.

1.3 Ergebnisse

1.3.1 Keine Evidenz für eine allgemeine Anreicherung von DNMs in nsCL/P

Der endgültige Datensatz enthielt 211 nsCL/P-Trios, 284 Kontroll-Trios und insgesamt 31.490 autosomale DNMs (Verteilung der DNMs, Indels und SNVs in Tabelle 1 der Publikation). Die durchschnittliche Anzahl von 63,6 autosomalen DNMs pro Trio lag im zu erwartenden Bereich (Kong et al., 2012) und zeigte keinen signifikanten Unterschied zwischen den Kohorten (64,1 vs. 63,3; P = 0,47). Auch die Untersuchung der verschiedenen Verteilungen in genomischen Regionen oder bezüglich einzelner Prädiktions-Scores ergab keine signifikanten Unterschiede (siehe Abbildung 1 der Publikation). Bei spezifischer Betrachtung der nsCL/P-DNMs mit besonders hohen CADD-Scores (≥ 30, n = 19) zeigten sich einzelne DNMs in Genen wie *WNT4*, *ALPI*, und *MYO10*, für die bereits eine Funktion in kraniofazialen Prozessen beschrieben wurde (Bachg et al., 2019; Iyyanar und Nazarali, 2017; Warner et al., 2009).

1.3.2 Anreicherung von DNMs in zwei TADs von GWAS-Regionen

In keiner Gruppe von untersuchten genomischen Elementen, die aufgrund von Konservierung (CNEs) oder gewebespezifischer Aktivität (Datensätze mit *Chromatin States*, VISTA Enhancer) relevant für die embryonale Gesichtsentwicklung sein könnten, konnte eine signifikante Anreicherung nach Korrektur für multiples Testen identifiziert werden, wobei die Power dieser Analyse durch die geringe Zahl an Beobachtungen eingeschränkt war (CNEs: 15 nsCL/P-DNMs vs. 19 Kontroll-DNMs, P = 0,88; VISTA-Enhancer: 14 nsCL/P-DNMs vs. 26 Kontroll-DNMs; P = 0,31). Auch bei Gruppierung der Enhancer nach gewebespezifischer Aktivität konnte kein signifikantes DNM-Enrichment nachgewiesen werden. Einzig die Untersuchung des DNM-Enrichments in den verschiedenen *Chromatin States* konnte eine nominal signifikante Anreicherung von nsCL/P-DNMs in bivalenten Transkriptionsstart-Sites und bivalenten Enhancern aus kraniofazialem Gewebe des Carnegie-Stadiums 15 zeigen (P = 0,03). Zusätzlich zeigte sich bei der Betrachtung des Gesamtmusters der Verteilung der nsCL/P-DNMs auf genomische Regionen in den verschiedenen *Chromatin States* ein interessantes, wenn auch nicht signifikantes Bild mit einer allgemeinen Überrepräsentation von nsCL/P-DNMs

in Enhancern, sowie einer Verminderung von nsCL/P in genomischen Regionen, die sich in nicht-aktiven Zuständen befinden (Abbildung 2A der Publikation).

Von den 2.991 untersuchten TADs zeigten 174 eine differenzielle Verteilung von DNMs zwischen den Kohorten, jedoch waren die Signifikanzniveaus nicht robust gegenüber multiplem Testen. In der Untergruppe der $TADs_{GWAS}$ zeigten sich zwei Loci mit starker Anreicherung von DNMs in der nsCL/P Kohorte: Am Locus 4q28 wurden 7 DNMs in nsCL/P, und keine in Kontrollen, beobachtet ($P = 8\times10^{-4}$). Am Locus 2p21 war die Verteilung 8 : 2 DNMs (P = 0,02). Am Locus 2p21 zeigte sich bei genauerer Betrachtung der Lokalisation außerdem eine Konzentration der nsCL/P-DNMs um die Region der GWAS-Varianten herum (Abbildung 2C der Publikation). Insgesamt konnten wir durch die Enrichment-Analyse in diesem TAD-Datensatz zwei Risikoloci aufzeigen, bei denen häufige und seltene Varianten vermutlich den gleichen biologischen Stoffwechselweg beeinflussen.

1.3.3 Identifizierung von Musculin als relevantem Transkriptionsfaktor für nsCL/P

Die Auswertung der *in silico* TFBS ergab sieben TF-Motive, die aufgrund verschiedener Evidenz Kandidaten für eine weitere Aufarbeitung darstellten (Tabelle S29 im *Supplement* der Publikation; JDP2 (var. 2), MSC, MEF2A, MAF::NFE, ATF3, SRF und NFE2L1). Durch Integration der *single-cell* Datensätze konnte schließlich weitere Evidenz dafür erhalten werden, dass MSC ein für die Ätiologie der nsCL/P relevanter TF sein könnte: MSC war u. a. an Embryonaltag 11.5 in Myozyten exprimiert (siehe Abbildung 3D der Publikation) und zeigte in den spezifischeren Expressionsdaten zur *lambdoidal junction* eine Expression in den Zellclustern *palatal epithelium* (Gaumenepithel) und *anterior and medial maxillary prominence* (Oberkiefervorsprünge) (Abbildung 3E der Publikation). Daraufhin wurde MSC für Validierungsexperimente ausgewählt. In Abbildung 4 ist hier für eine MSC-Bindestelle das Bindungsverhalten für Ref- bzw. Alt-Allel dargestellt. Die Ergebnisse bestätigten die von der computerbasierten Untersuchung vorhergesagten Effekte der DNMs.

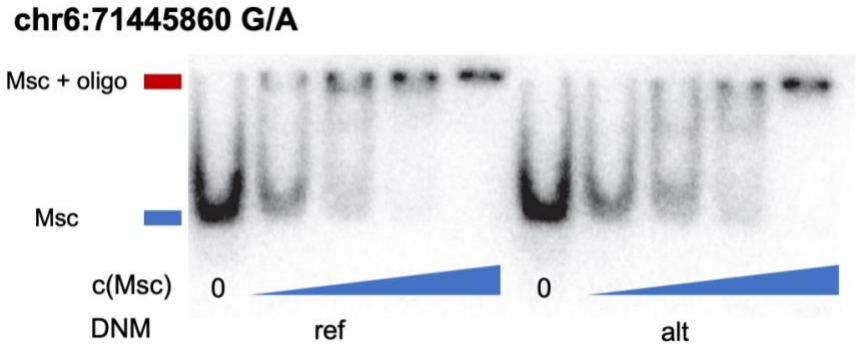

Abb. 4: *In-vitro*-Bindung von Musculin an DNM-Region

Dargestellt sind die Versuchsergebnisse für eine genomische Region mit *in silico* Bindungsstelle für Musculin (MSC). Die Bindungsaffinität und ihre Beeinflussung durch die DNM wurde *in vitro* mittels *Electrophoretic Mobility Shift Assays* unter Verwendung von Oligonukleotiden für Referenz- (ref) und Alternativ- (alt) Allel überprüft. Die DNM an der TFBS ist als zugehörige SNV auf dem + Strang angegeben. Hier ist ein Beispiel mit starker negativer Bindungsänderung (*loss of binding*) aus den *in silico* Daten ausgewählt. Für jede potenzielle TFBS wurden fünf verschiedene MSC-Konzentrationen für ref (linke Lanes) bzw. alt (rechte Lanes) titriert. Das Auftreten der oberen Bande (Msc+oligo) bei steigender MSC-Konzentration spiegelt eine Verschiebung des Molekulargewichts wider, was auf eine *in vitro* Bindung von MSC an das Oligonukleotid hinweist. Dargestellt ist ein bisher noch nicht veröffentlichtes Replikat des Assays für die TFBS auf Chromosom (chr.) 6 in Abbildung S16A in Zieger et al. (2023).

Abkürzungen: alt - Alternativ; DNM - *De novo* Mutation; MSC - Musculin; ref - Referenz; SNV - *single nucleotide variant*; TFBS - Transkriptionsfaktorbindestelle

1.4 Diskussion

Das Ziel dieser Arbeit war es, den Beitrag von nicht-kodierenden DNMs zur Ätiologie von nsCL/P zu untersuchen. Die Ergebnisse dieses Projekts liefern einerseits Hinweise für Kandidaten-DNMs und potenziell zugrundeliegende funktionelle Mechanismen. Andererseits können die hier entwickelten Strategien auf WGS-Daten anderer Phänotypen (speziell Fehlbildungen) angewandt werden.

Der untersuchte Phänotyp nsCL/P eignet sich gut für die Untersuchung von DNMs bei einer kleinen WGS-Kohorte. Da die Entstehung von nsCL/P auf eine kurze Zeitspanne während der Embryonalentwicklung und ausschließlich auf sich entwickelnde Gesichtsstrukturen beschränkt ist, handelt es sich bei nsCL/P um einen Phänotyp mit reduzierter biologischer Komplexität. Man kann daher annehmen, dass Risikovarianten

für nsCL/P, einschließlich DNMs, eine höhere Effektstärke aufweisen als bei anderen komplexeren multifaktoriellen Krankheiten. Unter Zuhilfenahme verfügbarer funktioneller Daten, die einen Teil der Biologie des Phänotyps abbilden, konnten hier interessante Ergebnisse erzielt werden.

Wir konnten einzelne Kandidaten-DNMs identifizieren, beispielsweise DNMs mit hohen CADD-Werten in vor allem exonischen Bereichen von Genen, die in der kraniofazialen Entwicklung eine Rolle spielen (siehe S. 15). Weiter befanden sich einige DNMs, die in VISTA-Enhancern oder CNEs lokalisiert waren, in Introns von Genen wie *FOXP2*, *MEIS2* und *HOXC5*, welche bereits als relevant für die Gaumenentwicklung beschrieben wurden (Cesario et al., 2016; Hirata et al., 2016; Machon et al., 2015). Diese Befunde legen nahe, weitere funktionelle Untersuchungen zum Beitrag nicht-kodierender Varianten in diesen Genen durchzuführen. Neben diesen isolierten Kandidaten-DNMs zeigte die Enrichment-Analyse der Chromatin-Daten Hinweise auf eine Anreicherung von nsCL/P-DNMs in aktiven Regionen in relevanten Gewebetypen, was sich bereits für häufige Varianten gezeigt hatte (Ludwig et al., 2017; Welzenbach et al., 2021). In diesem Zusammenhang sind auch die Ergebnisse der Enrichment-Analyse in den $TADs_{GWAS}$ als spannend zu bewerten: hier zeigte sich eine Konvergenz von seltenen und häufigen Varianten an zwei Loci. Dieser Fund deutet darauf hin, dass an den zwei identifizierten Loci häufige Varianten und seltene DNMs über einen gemeinsamen biologischen Stoffwechselweg wirken könnten. Beispielsweise beeinflussen die DNMs, die nahe der GWAS-Varianten des 2p21 Locus lokalisiert sind, vermutlich die Expression von *PKDCC*, welches ein Kandidatengen an diesem Locus ist (Ludwig et al., 2017). Eine solche Konvergenz konnte zuvor bereits für andere polygene Erkrankungen wie beispielsweise Autismus (Won et al., 2022) und Diabetes (Jurgens et al., 2022) gezeigt werden, für die nsCL/P bisher jedoch noch nicht.

Ein besonders zielführender Ansatz war die Verwendung von TFBS-Informationen, um DNMs von verschiedenen Loci durch ihre Position in einem molekularen Netzwerk zu verbinden. Der durch diesen Ansatz identifizierte Kandidaten-TF MSC deutet auf eine bisher nicht bekannte Beteiligung der sich entwickelnden Gesichtsmuskulatur in der nsCL/P-Ätiologie hin. Dieser Befund wird durch ältere Erkenntnisse gestützt: Der *Musculus orbicularis oris*, ein den Mund umschließender Gesichtsmuskel des Menschen, ist nachweislich auch bei nicht-betroffenen Verwandten von Individuen mit nsCL/P

verändert (Martin et al., 2000), was als subklinische Ausprägung von nsCL/P angesehen wird (Neiswanger et al., 2007). Auch das molekulare Netzwerk von Genen, welche im Rahmen der Gesichtsentwicklung miteinander interagieren, unterstützt die Hypothese (Abbildung 5): Neben MSC sind dort mehrere TFs beteiligt, die bereits entweder durch häufige Varianten in GWAS mit nsCL/P assoziiert wurden oder syndromalen Formen unterliegen (z. B. *TBX1* (Basha et al., 2018; Jugessur et al., 2009)).

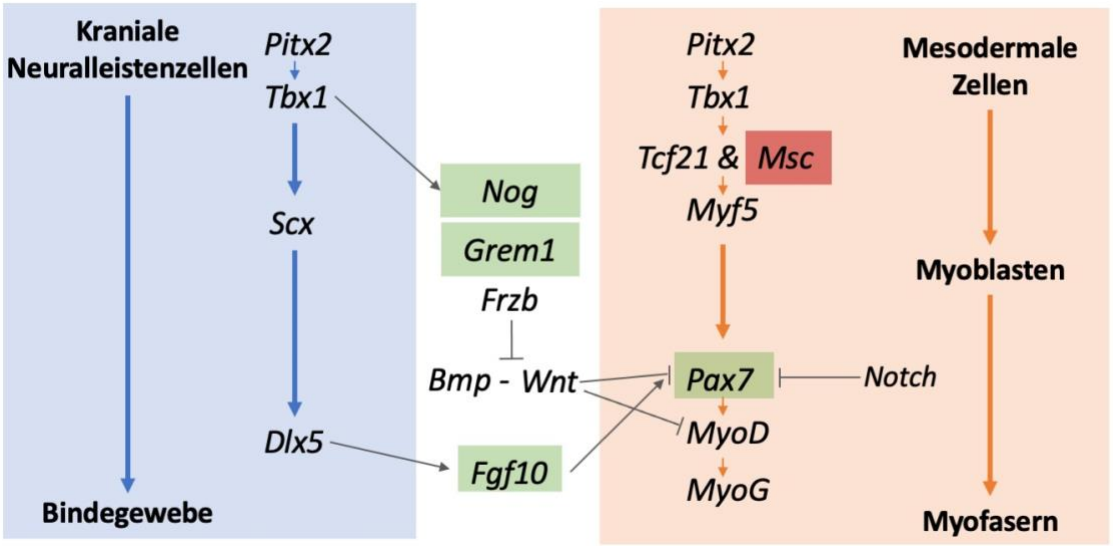

Abb. 5: Schematische Darstellung der Transkriptionsfaktoren, die an der embryonalen Entwicklung der Gesichtsmuskulatur beteiligt sind.

Hier ist das Netzwerk der Interaktion zwischen kranialen Neuralleistenzellen (blau hinterlegt) und mesodermalen Zellen, die sich über Myoblasten zu Muskelfasern entwickeln (orange hinterlegt) veranschaulicht. Gene, die in genomweiten Assoziationsstudien als Kandidatengene für nicht-syndromale Lippenspalten mit/ohne Gaumenspalten (nsCL/P) identifiziert wurden, sind grün markiert. Der Transkriptionsfaktor Musculin (MSC) mit Transkriptionsfaktorbindestellen in Regionen mit *de novo* Mutationen ist in rot hervorgehoben. Diese Abbildung entspricht einer deutschen Version der Figure S17 im *Supplement* von Zieger et al. (2023).

Die in diesem Projekt erhaltenen Ergebnisse unterliegen einigen Limitationen. Zunächst ist der verwendete Kontrolldatensatz zu beleuchten, welcher ebenfalls einer Krankheitskohorte zuzuordnen ist und v. a. angesichts der fehlenden Verfügbarkeit einer breiten Menge öffentlich zugänglicher WGS-Triodaten, insbesondere dem Fehlen eines Datensatzes mit phänotypisch gesunden Trios, ausgewählt wurde. Durch den gleichen Datenursprung der beiden Datensätze weisen sie jedoch ein ähnliches Studiendesign auf

und wurden mittels gleicher Sequenzier- und Auswertetechnologie untersucht. Dadurch konnte ein Großteil an Artefakten und Verzerrungen in unseren Analysen ausgeschlossen werden. Aufgrund epidemiologischer Hinweise für gemeinsame genetische Faktoren für die Entstehung von Krebs und orofazialen Spalten, die bislang jedoch nicht auf molekularer Ebene bestätigt wurden, sind unter Umständen pathogene DNMs verpasst worden (Zieger et al., 2023). Dadurch birgt die Kontroll-Kohorte jedoch insbesondere das Risiko falsch-negativer Ergebnisse, sodass in erster Linie kausale DNMs verpasst, jedoch keine falsch-positiven Befunde berichtet wurden (siehe *Supplemental Methods* S. 24 f.).

Die Interpretation von Varianten in nicht-kodierenden Bereichen des Genoms stellt immer noch eine Herausforderung für die genetische Forschung dar. In den hier durchgeführten Analysen scheinen auch die für nicht-kodierende SNVs trainierten *in silico* Prädiktions-Scores weiterhin insbesondere kodierende Effekte zu gewichten (bspw. fast ausschließlich exonische DNMs mit hohen CADD-Scores ≥ 30 (Tabelle S9 im *Supplement* der Publikation)). Auch ist die Bewertung der DNMs durch solche Scores bisher unabhängig von untersuchtem Phänotyp und relevantem Gewebetyp (Schipper et al., 2022), sodass eine weitere Verbesserung von *in silico* Prädiktions-Scores für nicht-kodierende Varianten notwendig ist. Die kodierenden DNMs des Datensatzes wurden anhand der *in silico* Scores und mit Hilfe des *Variant Effect Predictor* (McLaren et al., 2016) bewertet. Dies wird hier jedoch aufgrund des Schwerpunkts auf nicht-kodierende Varianten nicht separat dargestellt. Die identifizierten Kandidaten-DNMs werden aktuell jedoch durch das Bonner Team in der hiesigen Kohorte von nsCL/P Patienten resequenziert.

Ein letzter Aspekt sind mögliche Einschränkungen bezüglich der Berechnung der TFBS allgemein und speziell bei der Identifikation des Transkriptionsfaktors MSC. Generell ist (i) die Anwendung von PWMs und Konsensussequenzen zur Berechnung von TFBS eine Vereinfachung der Faktoren für eine Bindungsentstehung, weiter ist auch (ii) die Validierung der identifizierten TFBS durch EMSA nur eine *in vitro* Betrachtung der Bindung von Protein an eine genomische Sequenz. Dennoch konnten wir durch die EMSA die Bindung von MSC an die genomischen Sequenzen und auch den Einfluss der DNM bestätigen. Durch die weiteren unabhängigen Hinweise für eine Beteiligung der Gesichtsmuskulatur (s. o.) scheinen die Ergebnisse plausibel und eröffnen neue

Forschungsansätze zur Funktion von MSC und Muskelzellen für nsCL/P, z. B. durch *in vivo* Experimente in relevanten Zellsystemen.

Die Resultate dieses Promotionsprojekts bieten allerdings nicht nur Ansatzpunkte für anschließende Projekte zu Muskelzellsystemen (s. o.), sondern auch für die Erweiterung der Analysen des WGS-Datensatzes: Bisher wurden nur autosomale DNMs in die Analysen inkludiert; eine Auswertung der gonosomalen DNMs könnte weitere Erkenntnisse liefern. Zudem könnten die Individuen anderer Ethnien aus dem GMKF Projekt ausgewertet werden. Außerdem könnten als weiteres Projekt statt des relativ breit gefassten nsCL/P Phänotyps einzelne Subphänotypen betrachtet werden. Zwar wurden für dieses Projekt bereits Trios mit einem *cleft palate only* Phänotyp, bei dem nur der Gaumen betroffen ist, ausgeschlossen, da es Evidenz für eine getrennte Ätiologie dieser Erkrankungen gibt (Rahimov et al., 2012). Es erfolgte aber keine weitere Auftrennung nach der sogenannten LAHSHAL Klassifikation (Kriens, 1989), die eine genauere Unterscheidung des nsCL/P Phänotyps durch die Auftrennung der einzelnen CL/P Anteile in Lippenspalte (L), Kieferspalte (Alveolus = A), Hartgaumenspalte (H), und Segelspalte (S) ermöglicht. Eine solche Auftrennung wäre aufgrund von bereits bekannten Subphänotyp-Effekten bei häufigen Varianten (Ludwig et al., 2012; Ludwig et al., 2016) eine sinnvolle Erweiterung, andererseits würde dies die Anzahl an Trios pro Phänotyp und somit die statistische Power weiter reduzieren. Auch eine Weiterentwicklung des denovoLOBGOB Tools für die Auswertung direkt nebeneinander liegender SNVs und kleiner Indels sowie eine TF-Netzwerk-Analyse für die Gesamtheit der DNMs eines Individuums wären spannende Projekte. Weiter könnten identifizierte Kandidatenregionen mit interessanten DNMs in der Bonner Kohorte resequenziert werden.

Schlussendlich zeigt dieses Promotionsprojekt Methoden zur Priorisierung nicht-kodierender Varianten anhand funktioneller Daten zur Bewertung einzelner DNMs sowie zur Vernetzung von DNMs unterschiedlicher genomischer Lokalisationen auf, die auf Datensätze anderer Krankheitsbilder aber auch nsCL/P Datensätze wie die Bonner Kohorte angewandt werden können und hat mit der Entdeckung der potenziellen Rolle von MSC und dem TF-Netzwerk der fazialen Muskelentwicklung einen weiteren Gewebetyp, der bei der Ätiologie von nsCL/P mitwirken könnte, identifiziert.

1.5 Zusammenfassung

Im Rahmen dieses Promotionsprojektes wurde die Rolle von nicht-kodierenden *de novo* Mutationen (DNMs) bei der Entstehung nicht-syndromaler Lippen-Kiefer-Gaumenspalten (nsCL/P) untersucht. Dafür wurde ein Datensatz von Genomsequenzierungsdaten von betroffenen Individuen sowie deren Eltern aus dem *Gabriella Miller Kids First* Programm analysiert. Um ätiologisch relevante DNMs zu priorisieren, wurden die identifizierten DNMs u. a. mit funktionellen Datensätzen und Risiko-Scores annotiert und deren Verteilungen mit den Daten einer Kontroll-Kohorte verglichen. Während die meisten Vergleichsanalysen keine signifikanten Unterschiede zwischen den Kohorten zeigten, konnten wir signifikante Anreicherungen von DNMs in zwei *topologically associating domains* beobachten, welche bereits durch häufige Risikovarianten in der Ätiologie der nsCL/P (Locus 4q28 und 2p21) bekannt waren. Diese Konvergenz von häufigen und seltenen Varianten bestärkt die Relevanz dieser Risikoloci für nsCL/P und legt Varianten für mögliche funktionelle Untersuchungen nahe. Durch die systematische *in silico* Auswertung von Transkriptionsfaktorbindestellen und deren Beeinflussung durch DNMs wurden Kandidaten-Transkriptionsfaktoren vorhergesagt, welche über ihre molekularen Netzwerke zur nsCL/P beitragen könnten. Durch Integration von externen funktionellen Datensätzen und molekularen Validierungsexperimenten wurde dadurch der in der kraniofazialen Muskelentwicklung involvierte Transkriptionsfaktor Musculin (MSC) als Kandidat identifiziert. Eine mögliche Beteiligung von MSC in der Ätiologie von nsCL/P wird durch weitere subklinische Phänotypen von nsCL/P und das molekulare Netzwerk, in dem auch weitere, durch genomweite Assoziationsstudien identifizierte nsCL/P Risikogene involviert sind, bestätigt. Die Ergebnisse dieses Projekts liefern einerseits Hinweise für Kandidaten-DNMs und potenziell zugrundeliegende funktionelle Mechanismen für nsC/LP, andererseits sind die hier entwickelten Strategien auf WGS-Daten anderer Phänotypen (speziell Fehlbildungen) übertragbar.

1.6 Literaturverzeichnis der deutschen Zusammenfassung

Bachg AC, Horsthemke M, Skryabin BV, Klasen T, Nagelmann N, Faber C, Woodham E, Machesky LM, Bachg S, Stange R, Jeong HW, Adams RH, Bähler M, Hanley PJ. Phenotypic analysis of Myo10 knockout (Myo10tm2/tm2) mice lacking full-length (motorized) but not brain-specific headless myosin X. Sci Rep 2019; 9: 597

Basha M, Demeer B, Revencu N, Helaers R, Theys S, Bou Saba S, Boute O, Devauchelle B, Francois G, Bayet B, Vikkula M. 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 2018; 55: 449–458

Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, Jin S-C, Cooper ME, Dunnwald M, Mansilla MA, Leslie E, Bullard S, Lidral AC, Moreno LM, Menezes R, Vieira AR, Petrin A, Wilcox AJ, Lie RT, Jabs EW, Wu-Chou YH, Chen PK, Wang H, Ye X, Huang S, Yeow V, Chong SS, Jee SH, Shi B, Christensen K, Melbye M, Doheny KF, Pugh EW, Ling H, Castilla EE, Czeizel AE, Ma L, Field LL, Brody L, Pangilinan F, Mills JL, Molloy AM, Kirke PN, Scott JM, Arcos-Burgos M, Scott AF. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nat Genet 2010; 42: 525–529

Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida De Assis N, Alblas MA, Barth S, Freudenberg J, Lauster C, Schmidt G, Scheer M, Braumann B, Bergé SJ, Reich RH, Schiefke F, Hemprich A, Pötzsch S, Steegers-Theunissen RP, Pötzsch B, Moebus S, Horsthemke B, Kramer FJ, Wienker TF, Mossey PA, Propping P, Cichon S, Hoffmann P, Knapp M, Nöthen MM, Mangold E. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat Genet 2009; 41: 473–477

Buecker C, Wysocka J. Enhancers as information integration hubs in development: lessons from genomics. Trends Genet 2012; 28: 276–284

Bulger M, Groudine M. Enhancers: The abundance and function of regulatory sequences beyond promoters. Dev Biol 2010; 339: 250–257

Butali A, Mossey P, Adeyemo WL, Eshete M, Gaines LAL, Braimah RO, Aregbesola BS, Rigdon J, Emeka C, Olutayo J, Ogunlewe O, Ladeinde A, Abate F, Hailu T, Mohammed I, Gravem P, Deribew M, Gesses M, Adeyemo A, Marazita M, Murray J. Rare functional variants in genome-wide association identified candidate genes for nonsyndromic clefts in the African population. Am J Med Genet A 2014; 164: 2567–2571

Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J. The single-cell transcriptional landscape of mammalian organogenesis. Nature 2019; 566: 496–502

Cesario JM, Almaidhan AA, Jeong J. Expression of forkhead box transcription factor genes Foxp1 and Foxp2 during jaw development. Gene Expr Patterns 2016; 20: 111–119

Christensen K, Juel K, Herskind AM, Murray JC. Long term follow up study of survival associated with cleft lip and palate at birth. BMJ 2004; 328: 1405

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012; 485: 376–380

Fakhouri WD, Rahimov F, Attanasio C, Kouwenhoven EN, Ferreira De Lima, R. L., Felix TM, Nitschke L, Huver D, Barrons J, Kousa YA, Leslie E, Pennacchio LA, van Bokhoven H, Visel A, Zhou H, Murray JC, Schutte BC. An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. Hum Mol Genet 2014; 23: 2711–2720

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, Santana-Garcia W, Tan G, Chèneby J, Ballester B, Parcy F, Sandelin A, Lenhard B, Wasserman WW, Mathelier A. JASPAR

2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2020; 48: D87–D92

Grosen D, Bille C, Petersen I, Skytthe A, Hjelmborg JvB, Pedersen JK, Murray JC, Christensen K. Risk of oral clefts in twins. Epidemiology 2011; 22: 313–319

Grosen D, Chevrier C, Skytthe A, Bille C, Mølsted K, Sivertsen A, Murray JC, Christensen K, Sivertsen Å. A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. J Med Genet 2010; 47: 162–168

Hirata A, Katayama K, Tsuji T, Imura H, Natsume N, Sugahara T, Kunieda T, Nakamura H, Otsuki Y. Homeobox family Hoxc localization during murine palate formation. Congenit Anom (Kyoto) 2016; 56: 172–179

Iyyanar PP, Nazarali AJ. Hoxa2 inhibits Bone Morphogenetic Protein Signaling during Osteogenic Differentiation of the Palatal Mesenchyme. Front Physiol 2017; 8: 929

Jones MR, Peng PC, Coetzee SG, Tyrer J, Reyes ALP, Corona RI, Davis B, Chen S, Dezem F, Seo JH, Kar S, Dareng E, Berman BP, Freedman ML, Plummer JT, Lawrenson K, Pharoah P, Hazelett DJ, Gayther SA. Ovarian Cancer Risk Variants Are Enriched in Histotype-Specific Enhancers and Disrupt Transcription Factor Binding Sites. Am J Hum Genet 2020; 107: 622–635

Jugessur A, Farlie PG, Kilpatrick N. The genetics of isolated orofacial clefts: from genotypes to subphenotypes. Oral Dis 2009; 15: 437–453

Jugessur A, Murray JC. Orofacial clefting: recent insights into a complex trait. Curr Opin Genet Dev 2005; 15: 270–278

Jurgens SJ, Choi SH, Morrill VN, Chaffin M, Pirruccello JP, Halford JL, Weng LC, Nauffal V, Roselli C, Hall AW, Oetjens MT, Lagerman B, VanMaanen DP, Abecasis G, Bai X,

Balasubramanian S, Baras A, Beechert C, Boutkov B, Cantor M, Coppola G, De T, Deubler A, Economides A, Eom G, Ferreira MA, Forsythe C, Fuller ED, Gu Z, Habegger L, Hawes A, Jones MB, Karalis K, Khalid S, Krasheninina O, Lanche R, Lattari M, Li D, Lopez A, Lotta LA, Manoochehri K, Mansfield AJ, Maxwell EK, Mighty J, Mitnaul LJ, Nafde M, Nielsen J, O'Keeffe S, Orelus M, Overton JD, Padilla MS, Panea R, Polanco T, Pradhan M, Rasool A, Reid JG, Salerno W, Schleicher TD, Shuldiner A, Siminovitch K, Staples JC, Ulloa RH, Verweij N, Widom L, Wolf SE, Aragam KG, Lunetta KL, Haggerty CM, Lubitz SA, Ellinor PT. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nat Genet 2022; 54: 240–250

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database Consortium; Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020; 581: 434–443

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014; 46: 310–315

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk. Nature 2012; 488: 471–475

Kriens O. LAHSHAL - A concise documentation system for cleft lip, alveolus and palate diagnoses. In: Kriens O., Hrsg. What is a cleft lip and palate? Proceedings of an Advanced Workshop, Bremen 1987. Stuttgart: Georg Thieme Verlag, 1989: 139–142

Lansdon LA, Dickinson A, Arlis S, Liu H, Hlas A, Hahn A, Bonde G, Long A, Standley J, Tyryshkina A, Wehby G, Lee NR, Daack-Hirsch S, Mohlke K, Girirajan S, Darbro BW, Cornell RA, Houston DW, Murray JC, Manak JR. Genome-wide analysis of copy-number variation in humans with cleft lip and/or cleft palate identifies COBLL1, RIC1, and ARHGEF38 as clefting genes. Am J Hum Genet 2023; 110: 71–91

Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, Jain D, Laurie CC, Doheny KF, McHenry T, Resick J, Sanchez C, Jacobs J, Emanuele B, Vieira AR, Neiswanger K, Lidral AC, Valencia-Ramirez LC, Lopez-Palacio AM, Valencia DR, Arcos-Burgos M, Czeizel AE, Field LL, Padilla CD, Cutiongco-de la Paz, Eva Maria, C., Deleyiannis F, Christensen K, Munger RG, Lie RT, Wilcox A, Romitti PA, Castilla EE, Mereb JC, Poletta FA, Orioli IM, Carvalho FM, Hecht JT, Blanton SH, Buxó CJ, Butali A, Mossey PA, Adeyemo WL, James O, Braimah RO, Aregbesola BS, Eshete MA, Abate F, Koruyucu M, Seymen F, Ma L, Salamanca JE de, Weinberg SM, Moreno L, Murray JC, Marazita ML. A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24.2, 17q23 and 19q13. Hum Mol Genet 2016; 25: 2862–2872

Letra A, Maili L, Mulliken JB, Buchanan E, Blanton SH, Hecht JT. Further evidence suggesting a role for variation in ARHGAP29 variants in nonsyndromic cleft lip/palate. Birth Defects Res A Clin Mol Teratol 2014; 100: 679–685

Li H, Jones KL, Hooper JE, Williams T. The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. Development 2019; 146: dev174888

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive

mapping of long range interactions reveals folding principles of the human genome. Science 2009; 326: 289–293

Little J, Cardy A, Munger RG. Tobacco smoking and oral clefts: a meta-analysis. Bull World Health Organ 2004; 82: 213–218

Ludwig KU, Ahmed ST, Böhmer AC, Sangani NB, Varghese S, Klamt J, Schuenke H, Gültepe P, Hofmann A, Rubini M, Aldhorae KA, Steegers-Theunissen RP, Rojas-Martinez A, Reiter R, Borck G, Knapp M, Nakatomi M, Graf D, Mangold E, Peters H. 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 2016; 12: e1005914

Ludwig KU, Böhmer AC, Bowes J, Nikolić M, Ishorst N, Wyatt N, Hammond NL, Gölz L, Thieme F, Barth S, Schuenke H, Klamt J, Spielmann M, Aldhorae K, Rojas-Martinez A, Nöthen MM, Rada-Iglesias A, Dixon MJ, Knapp M, Mangold E. 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 2017; 26: 829–842

Ludwig KU, Degenhardt F, Nöthen MM. Die Rolle seltener Varianten bei häufigen Krankheiten. Med Genet 2019; 31: 212–221

Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E, Böhmer AC, Mattheisen M, Alblas MA, Barth S, Kluck N, Lauster C, Braumann B, Reich RH, Hemprich A, Pötzsch S, Blaumeiser B, Daratsianos N, Kreusch T, Murray JC, Marazita ML, Ruczinski I, Scott AF, Beaty TH, Kramer F-J, Wienker TF, Steegers-Theunissen RP, Rubini M, Mossey PA, Hoffmann P, Lange C, Cichon S, Propping P, Knapp M, Nöthen MM. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. Nat Genet 2012; 44: 968–971

Machon O, Masek J, Machonova O, Krauss S, Kozmik Z. Meis2 is essential for cranial and cardiac neural crest development. BMC Dev Biol 2015; 15: 40

Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, Assis NA de, Chawa TA, Mattheisen M, Steffens M, Barth S, Kluck N, Paul A, Becker J, Lauster C, Schmidt G, Braumann B, Scheer M, Reich RH, Hemprich A, Pötzsch S, Blaumeiser B, Moebus S, Krawczak M, Schreiber S, Meitinger T, Wichmann H-E, Steegers-Theunissen RP, Kramer F-J, Cichon S, Propping P, Wienker TF, Knapp M, Rubini M, Mossey PA, Hoffmann P, Nöthen MM. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. Nat Genet 2010; 42: 24–26

Mangold E, Ludwig KU, Nöthen MM. Breakthroughs in the genetics of orofacial clefting. Trends Mol Med 2011; 17: 725–733

Martin RA, Hunter V, Neufeld-Kaiser W, Flodman P, Spence MA, Furnas D, Martin KA. Ultrasonographic detection of orbicularis oris defects in first degree relatives of isolated cleft lip patients. Am J Med Genet 2000; 90: 155–161

McGuire AL, Caulfield T, Cho MK. Research ethics and the challenge of whole-genome sequencing. Nat Rev Genet 2008; 9: 152–156

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol 2016; 17: 122

Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. Lancet 2009; 374: 1773-85

Mossey PA, Modell B. Epidemiology of Oral Clefts 2012: An International Perspective. Front Oral Biol 2012; 16: 1–18

Neiswanger K, Weinberg SM, Rogers CR, Brandon CA, Cooper ME, Bardi KM, Deleyiannis FW, Resick JM, Bowen A, Mooney MP, Salamanca JE de, González B, Maher BS, Martin RA, Marazita ML. Orbicularis oris muscle defects as an expanded phenotypic feature in nonsyndromic cleft lip with or without cleft palate. Am J Med Genet A 2007; 143A: 1143–1149

Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell 2015; 163: 68–83

Rada-Iglesias A, Bajpai R, Prescott S, Brugmann SA, Swigut T, Wysocka J. Epigenomic Annotation of Enhancers Predicts Transcriptional Regulators of Human Neural Crest. Cell Stem Cell 2012; 11: 633–648

Rahimov F, Jugessur A, Murray JC. Genetics of nonsyndromic orofacial clefts. Cleft Palate Craniofac J 2012; 49: 73–91

Rainger JK, Bhatia S, Bengani H, Gautier P, Rainger J, Pearson M, Ansari M, Crow J, Mehendale F, Palinkasova B, Dixon MJ, Thompson PJ, Matarin M, Sisodiya SM, Kleinjan DA, Fitzpatrick DR. Disruption of SATB2 or its long-range cis-regulation by SOX9 causes a syndromic form of Pierre Robin sequence. Hum Mol Genet 2014; 23: 2569–2579

Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell 2011; 147: 57–69

Rosero Salazar DH, Carvajal Monroy PL, Wagener FADTG, Von den Hoff JW. Orofacial Muscles: Embryonic Development and Regeneration after Injury. J Dent Res 2020; 99: 125–132

Savastano CP, Brito LA, Faria C, Setó-Salvia N, Peskett E, Musso CM, Alvizi L, Ezquina SA, James C, GOSgene, Beales P, Lees M, Moore GE, Stanier P, Passos-Bueno MR. Impact of rare variants in ARHGAP29 to the etiology of oral clefts: role of loss-of-function vs missense variants. Clin Genet 2017; 91: 683–689

Schipper M, Posthuma D. Demystifying non-coding GWAS variants: an overview of computational tools and methods. Hum Mol Genet 2022; 31: R73–R83

Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, Wright CF, Firth HV, Fitzpatrick DR, Barrett JC, Hurles ME. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature 2018; 555: 611–616

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

Sylvester B, Brindopke F, Suzuki A, Giron M, Auslander A, Maas RL, Tsai B, Gao H, Magee W, Cox TC, Sanchez-Lara PA. A Synonymous Exonic Splice Silencer Variant in IRF6 as a Novel and Cryptic Cause of Non-Syndromic Cleft Lip and Palate. Genes (Basel) 2020; 11: 903

Thieme F, Ludwig KU. The Role of Noncoding Genetic Variation in Isolated Orofacial Clefts. J Dent Res 2017; 96: 1238–1247

Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 2014; 42: D975–D979

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nat Rev Methods Primers 2021; 1: 59

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser-a database of tissue-specific human enhancers. Nucleic Acids Res 2007; 35: D88–D92

Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 2008; 9: 255–266

Warner DR, Smith HS, Webb CL, Greene RM, Pisano MM. Expression of Wnts in the developing murine secondary palate. Int J Dev Biol 2009; 53: 1105–1112

Welzenbach J, Hammond NL, Nikolić M, Thieme F, Ishorst N, Leslie EJ, Weinberg SM, Beaty TH, Marazita ML, Mangold E, Knapp M, Cotney J, Rada-Iglesias A, Dixon MJ, Ludwig KU. Integrative approaches generate insights into the architecture of non-syndromic cleft lip ± cleft palate. HGG Adv 2021; 2: 100038

Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. Cell Rep 2018; 23: 1581–1597

Won H, Huguet G, Jacquemont S. Rare and common autism risk variants converge across 16p. Nat Genet 2022; 54: 1587–1588

Yu Y, Zuo X, He M, Gao J, Fu Y, Qin C, Meng L, Wang W, Song Y, Cheng Y, Zhou F, Chen G, Zheng X, Wang X, Liang B, Zhu Z, Fu X, Sheng Y, Hao J, Liu Z, Yan H, Mangold E, Ruczinski I, Liu J, Marazita ML, Ludwig KU, Beaty TH, Zhang X, Sun L, Bian Z. Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat Commun 2017; 8: 14364

Zieger HK, Weinhold L, Schmidt A, Holtgrewe M, Juranek SA, Siewert A, Scheer AB, Thieme F, Mangold E, Ishorst N, Brand FU, Welzenbach J, Beule D, Paeschke K, Krawitz PM, Ludwig KU. Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of de novo mutations. HGG Adv 2023; 4: 100166

2. Veröffentlichung

HGG Advances

ARTICLE

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

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

Summary

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

Introduction

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

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

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

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

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

Material and methods

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

Subjects and data resources

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

WGS data analysis and variant calling

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

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

Analysis of DNM enrichment in genomic features

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

Analysis of topologically associating domains

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

Analysis of DNMs in TF binding sites

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

Single-cell expression data

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

Electrophoretic mobility shift assays

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

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

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

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

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

Results

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

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

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

Identification of deleterious variants in craniofacial genes

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

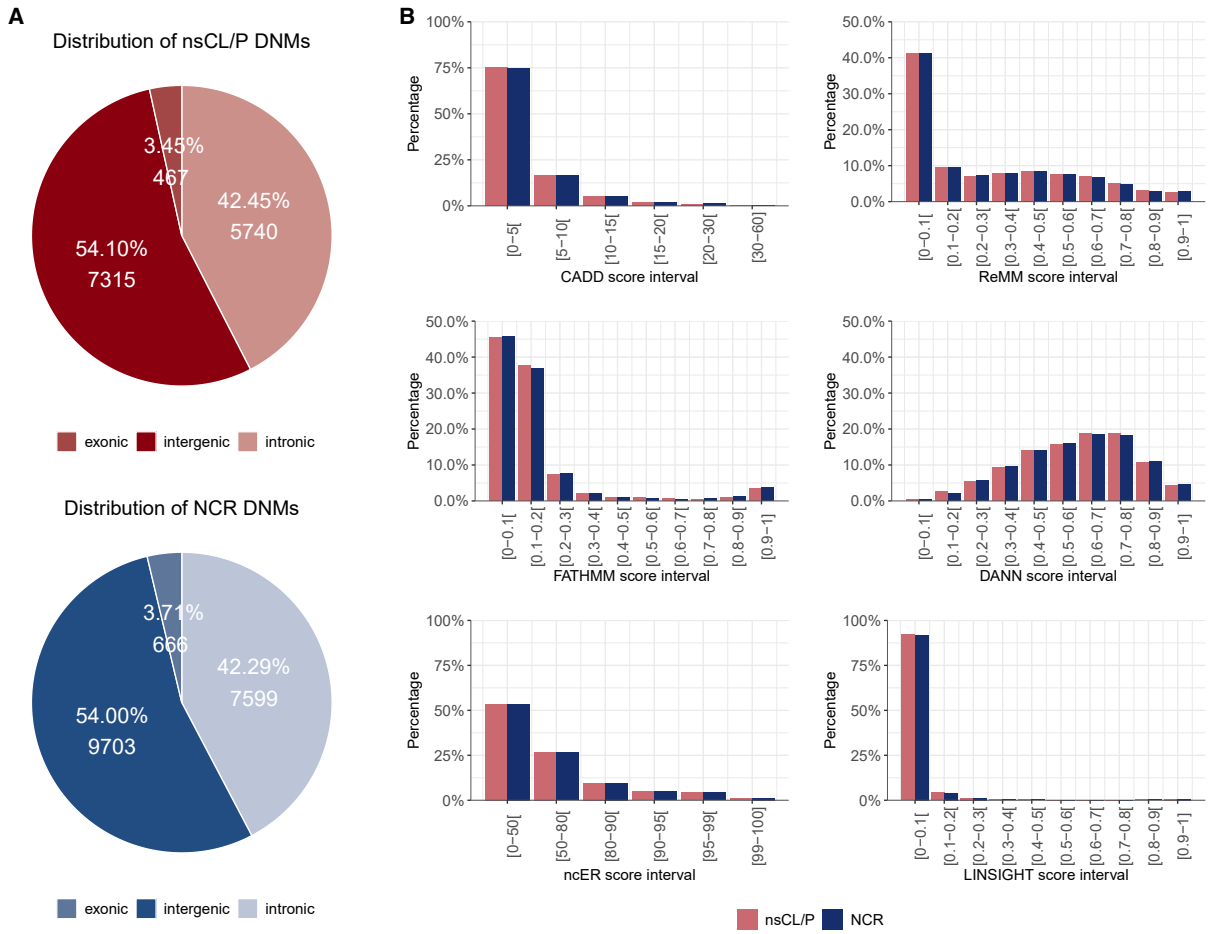

Figure 1. Comparative analyses of *de novo* mutations
(A) *De novo* mutations (DNMs) observed in non-syndromic cleft lip with/without cleft palate (nsCL/P) case-parent trios (red) and NCR trios (blue) were annotated according to genomic location (i.e., exonic/intronic/intergenic). Exonic DNMs were defined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

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

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

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

Convergence of non-coding DNMs at two GWAS risk loci

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

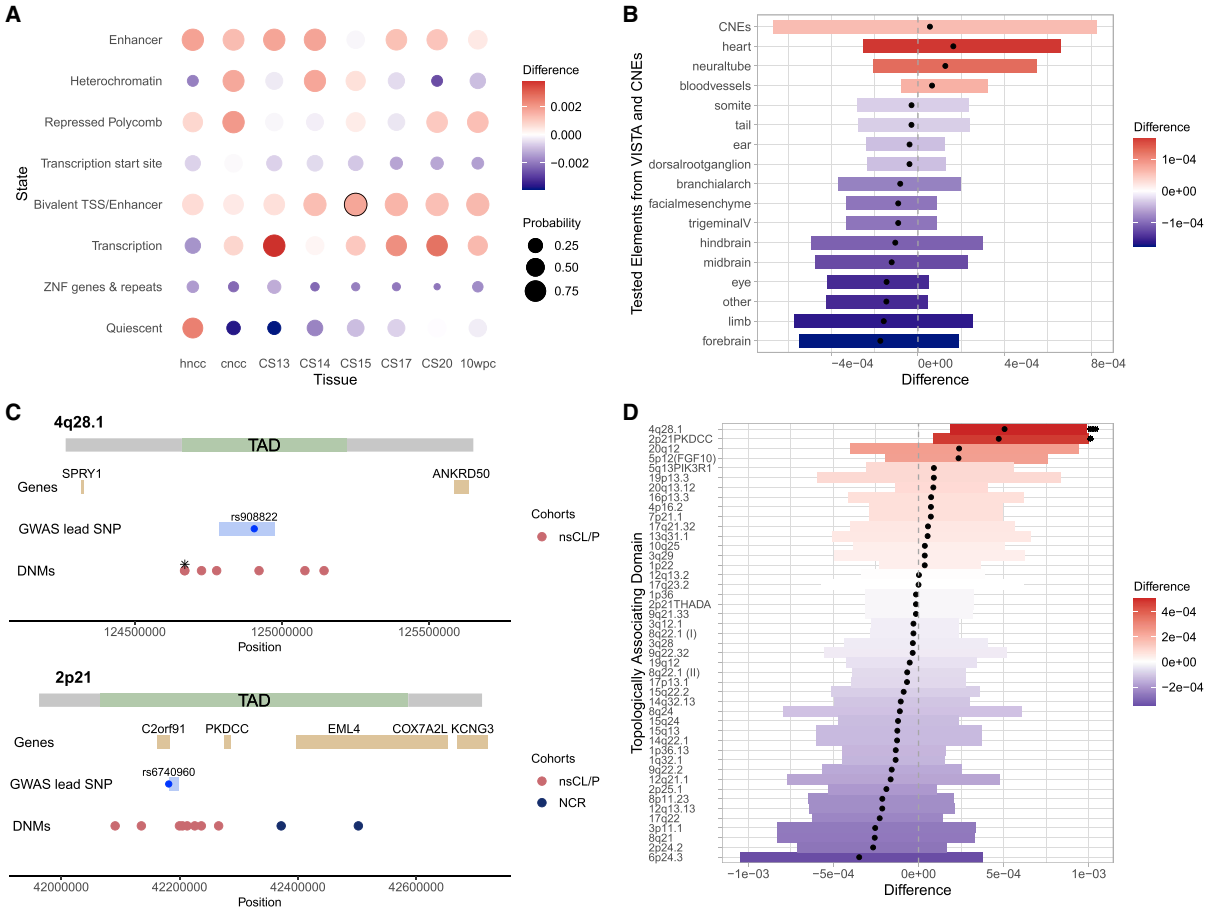

Figure 2. Enrichment of non-syndromic cleft lip with/without cleft palate *de novo* mutations in genomic candidate regions
(A) DNMs were mapped in eight chromatin states derived from human neural crest cells (hNCCs), cranial neural crest cells (cNCCs), and human embryonic facial tissue. FunciVar enrichment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

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

Identification of candidate TFs

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

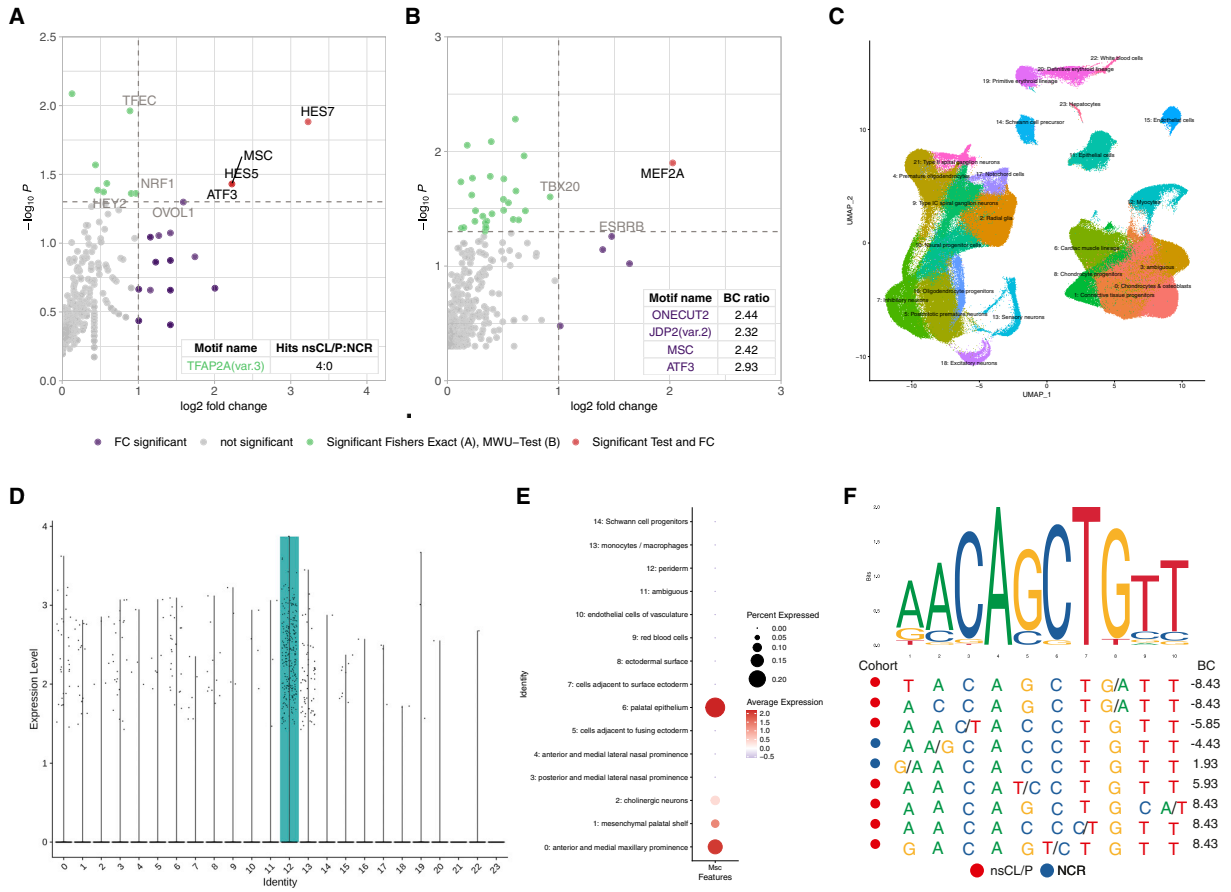

Figure 3. Identification of Musculin as a player in non-syndromic cleft lip with/without cleft palate etiology

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

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

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

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

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

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

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

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

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

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

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

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

DNMs in MSC binding sites affect binding *in vitro*

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

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

Discussion

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

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

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

Second, despite the limited evidence for overall enrichments, we identified a convergence of DNMs at loci that had prior evidence for an involvement in nsCL/P. Most interestingly, we observed a significant overrepresentation of DNMs in regions that were previously implicated in nsCL/P etiology by common variants. Specifically, two risk loci, 4q28 and $2p21_{PKDCC}$, harbored significantly more DNMs in nsCL/P trios than the reference cohort. At 2p21, the variants clustered within a region of 175 kb, in close vicinity to rs6740960, which has been suggested as the sole causal variant at this locus.[39,46] As another example, we observed two intronic DNMs in the nsCL/P candidate gene, *ZFHX4*,[11] for which a frameshift mutation was previously reported (Table S31). While the exact functional effect and molecular mechanisms of these non-coding DNMs at GWAS loci or within candidate gene loci remain unclear, these findings illustrate the presence of allelic heterogeneity at established loci and pave the way for functional follow-up studies.

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

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

Data and code availability

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

Supplemental information

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

Acknowledgments

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

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

Author contributions

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

Declaration of interests

The authors declare no competing interests.

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

Web resources

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

3. Danksagung

Mein besonderer Dank gilt meiner Doktormutter Prof. Kerstin Ludwig, die mich über all die Jahre in diesem Projekt unterstützt und durch die Journal-Veröffentlichung dieser Promotion begleitet hat. Vielen herzlichen Dank für die unglaubliche Förderung, die Betreuung mit zahlreichen wissenschaftlichen und persönlichen Gesprächen, das Herausfordern und das Bestärken.

Mein weiterer Dank gilt all den Personen, die dieses große Projekt möglich gemacht haben: Dr. Leonie Weinhold für die vielen statistischen Gespräche und die Geduld. Weiter danken möchte ich Prof. Katrin Paeschke und Dr. Stefan Juranek für das Ermöglichen der EMSA-Versuche im Labor. Zudem möchte ich Dr. Axel Schmidt, Frederic Thieme, Anna Siewert, Dr. Julia Heggemann (geb. Welzenbach), Dr. Nina Ishorst und PD Dr. Elisabeth Mangold für die tolle Zusammenarbeit danken. Mein besonderer Dank gilt weiter den Kooperationspartnern Prof. Dieter Beule und Dr. Manuel Holtgrewe in Berlin sowie Prof. Peter Krawitz und Fabian Brand in Bonn, die die Analysen eines so großen Datensatzes ermöglicht haben.

Außerdem danke ich meiner Familie für den Beistand über all die Jahre, für das Zuhören, Gegenlesen, und unzählige Gespräche. Danke für die beruhigenden Worte. Ich liebe euch!

4. Publikationen und Kongressbeiträge

Publikationen

Zieger HK, Weinhold L, Schmidt A, Holtgrewe M, Juranek SA, Siewert A, Scheer AB, Thieme F, Mangold E, Ishorst N, Brand FU, Welzenbach J, Beule D, Paeschke K, Krawitz PM, Ludwig KU. Prioritization of non-coding elements involved in non-syndromic cleft lip with/without cleft palate through genome-wide analysis of *de novo* mutations. HGG Adv. 2022 Dec 5;4(1):100166. doi: 10.1016/j.xhgg.2022.100166. PMID: 3658941

Thieme F, Henschel L, Hammond NL, Ishorst N, Hausen J, Adamson AD, Biedermann A, Bowes J, Zieger HK, Maj C, Kruse T, Buness A, Hoischen A, Gilissen C, Kreusch T, Jäger A, Gölz L, Braumann B, Aldhorae K, Rojas-Martinez A, Krawitz PM, Mangold E, Dixon MJ, Ludwig KU. Extending the allelic spectrum at noncoding risk loci of orofacial clefting. Hum Mutat. 2021 Aug;42(8):1066-1078. doi: 10.1002/humu.24219. Epub 2021 Jun 3. PMID: 34004033.

Kongressbeiträge

ESHG 06/2020, Zieger HK, Thieme F, Schmidt A, Fazaal J, Mangold E, Welzenbach J, Ludwig KU: “Using transcription factor sequence information to interpret rare variants in non-coding regions”, P17.114.A

ESHG 08/2021, Zieger HK, Weinhold L, Schmidt A, Holtgrewe H, Juranek SA, Siewert A, Thieme F, Brand F, Welzenbach J, Beule D, Paeschke K, Krawitz PM, Ludwig KU: “Systematic analysis of non-coding de novo mutations from whole genome sequence data of triads with non-syndromic cleft lip with/without cleft palate”, P04.064.A

GfH Jahrestagung 03/2022, Zieger HK, Weinhold L, Schmidt A, Holtgrewe H, Juranek SA, Siewert A, Thieme F, Brand F, Welzenbach J, Beule D, Paeschke K, Krawitz PM, Ludwig KU: “Systematic analysis of non-coding *de novo* mutations from whole genome sequence data of triads with non-syndromic cleft lip with/without cleft palate”, SEL-002