Extending the allelic spectrum at noncoding risk loci of orofacial clefting

DOI:
10.1002/humu.24219

Citation for published version (APA):
Thieme, F., Henschel, L., Hammond, N. L., Ishorst, N., Hausen, J., Adamson, A. D., Biedermann, A., Bowes, J., Zieger, H. K., Maj, C., Kruse, T., Buness, A., Hoischen, A., Gilissen, C., Kreusch, T., Jäger, A., Gölz, L., Braumann, B., Aldhorae, K., ... Ludwig, K. U. (2021). Extending the allelic spectrum at noncoding risk loci of orofacial clefting. *Human Mutation*, *42*(8), 1066-1078. https://doi.org/10.1002/humu.24219

Published in:
Human Mutation

Received: 12 March 2020 | Revised: 3 May 2021 | Accepted: 15 May 2021

DOI: 10.1002/humu.24219

METHODS

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Extending the allelic spectrum at noncoding risk loci of orofacial clefting

Frederic Thieme[1] | Leonie Henschel[1] | Nigel L. Hammond[2] | Nina Ishorst[1] | Jonas Hausen[3,4] | Antony D. Adamson[2] | Angelika Biedermann[1] | John Bowes[5] | Hanna K. Zieger[1] | Carlo Maj[3] | Teresa Kruse[6] | Andreas Buness[3,4] | Alexander Hoischen[7,8,9] | Christian Gilissen[7,9] | Thomas Kreusch[10] | Andreas Jäger[11] | Lina Gölz[11,12] | Bert Braumann[6] | Khalid Aldhorae[13] | Augusto Rojas-Martinez[14] | Peter M. Krawitz[3] | Elisabeth Mangold[1] | Michael J. Dixon[2] | Kerstin U. Ludwig[1]

[1]Institute of Human Genetics, School of Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany

[2]Faculty of Biology, Medicine, and Health, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK

[3]School of Medicine, Institute of Genomic Statistics and Bioinformatics, University Hospital Bonn, University of Bonn, Bonn, Germany

[4]Department of Medical Biometry, Informatics, and Epidemiology, School of Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany

[5]Arthritis Research UK Centre for Genetics and Genomics, University of Manchester, Manchester, UK

[6]Department of Orthodontics, University of Cologne, Cologne, Germany

[7]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands

[8]Department of Internal Medicine, Radboud University Medical Center, Nijmegen, The Netherlands

[9]Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands

[10]Department of Oral and Maxillofacial Surgery, Head and Neck Centre, Asklepios Klinik Nord, Heidberg, Hamburg, Germany

[11]Department of Orthodontics, University of Bonn, Bonn, Germany

[12]Department of Orthodontics, University of Erlangen-Nuremberg, Erlangen, Germany

[13]Department of Orthodontics, Thamar University, Thamar, Yemen

[14]Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, and Universidad Autonoma de Nuevo Leon, Centro de Investigación y Desarrollo en Ciencias de la Salud, Monterrey, Mexico

Correspondence

Kerstin U. Ludwig, Venusberg-Campus 1, building 76, 53127 Bonn, Germany.
Email: kludwig1@uni-bonn.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: LU 1944-3/1

Abstract

Genome-wide association studies (GWAS) have generated unprecedented insights into the genetic etiology of orofacial clefting (OFC). The moderate effect sizes of associated noncoding risk variants and limited access to disease-relevant tissue represent considerable challenges for biological interpretation of genetic findings. As rare variants with stronger effect sizes are likely to also contribute to OFC, an alternative approach to delineate pathogenic mechanisms is to identify private mutations and/or an increased burden of rare variants in associated regions. This report describes a framework for targeted resequencing at selected noncoding risk loci contributing to nonsyndromic cleft lip with/without cleft palate (nsCL/P), the most frequent OFC subtype. Based on GWAS data, we selected three risk loci and identified candidate regulatory regions (CRRs) through the integration of credible SNP information, epigenetic data from relevant cells/tissues, and conservation scores. The CRRs (total 57 kb) were resequenced in a multiethnic study population (1061 patients; 1591 controls), using single-molecule molecular inversion probe technology. Combining evidence from in silico variant annotation, pedigree- and burden analyses, we identified 16 likely deleterious rare variants that represent new candidates for functional studies in nsCL/P. Our framework is scalable and represents a promising approach to the investigation of additional congenital malformations with multifactorial etiology.

KEYWORDS

cleft palate, connective tissue biology, craniofacial anomalies, craniofacial biology/genetics, developmental biology, epidemiology, genetics, genomics, orofacial cleft(s)

1 | INTRODUCTION

Orofacial clefting ranks among the most common birth defects and imposes a substantial burden on affected individuals and their families. The most frequent subtype is nonsyndromic cleft lip with or without cleft palate (nsCL/P), with a prevalence of around 1 in 1000 live births (Mangold et al., 2011). NsCL/P has a multifactorial etiology, whereby both environmental and genetic factors contribute to disease risk. Importantly, genetic factors play a major role in nsCL/P, with heritability estimates being reported as high as 90% in twins (Grosen et al., 2011).

Recent systematic approaches, such as genome-wide association studies (GWAS) and meta-analyses thereof, have yielded unprecedented insights into the genetic etiology of many multifactorial diseases. For nsCL/P, these analyses have led to the identification of around 40 risk loci (e.g., Leslie et al., 2017; Ludwig et al., 2017; Thieme and Ludwig, 2017; Yu et al., 2017). As for many other common diseases (Maurano et al., 2012), these studies have revealed that the common nsCL/P risk variants: (i) are mainly located in noncoding regions of the human genome; (ii) have low to moderate effect sizes; and (iii) are often highly correlated with one another (as reflected through strong linkage disequilibrium (LD; Thieme and Ludwig, 2017). These characteristics render the inference of causal variants and/or causal regulatory mechanisms challenging. For many multifactorial disorders, approaches such as expression quantitative trait loci analyses, whereby genetic risk variants are correlated with the gene expression levels, have proven successful (Gamazon et al., 2018). However, this methodology requires the availability of disease-relevant cell types and tissues from multiple donors. In nsCL/P, access to relevant human tissue is limited due to the embryonic time point of face formation. Therefore, to date, identification of the functional effects of common nsCL/P risk variants has been achieved for only a limited number of nsCL/P risk loci (Leslie et al., 2015; Liu et al., 2017; Rahimov et al., 2008).

Increasing evidence suggests that rare variants play a role in the etiology of nsCL/P. This includes the observation of stable prevalence rates, despite strong evolutionary negative selection (Mossey and Modell, 2012) and the lack of explained heritability after accounting for common risk variants (Ludwig et al., 2017). Rare variants can be located in yet undescribed genomic regions, where their identification might suggest novel risk loci that contribute to the genetic architecture of nsCL/P. However, rare variants may also be found in previously identified risk loci characterized by common variants (Rivas et al., 2011). At these loci, rare variants can facilitate follow-up analyses based on (i) the absence of strong LD, and (ii) the presumably larger effect sizes. So far, the majority of rare variant analyses in nsCL/P have focused on the protein-coding regions of the human genome, using approaches, such as whole-exome (Basha et al., 2018; Cox et al., 2018) and candidate gene (Marini et al., 2019; Savastano et al., 2016) sequencing. For noncoding regions, available results for rare variants are largely limited to those generated from a small number of individual families (Cvjetkovic et al., 2015; Fakhouri et al., 2014). However, two recent studies have shed light on the role of low-frequency, private noncoding mutations in large study populations of nsCL/P patients. In 2015, Leslie et al. (2015) resequenced a selected set of GWAS/linkage loci in a trio-based cohort and identified de novo variants in 8% of the patient study population. For one variant, an allele-specific effect on enhancer activity was demonstrated. In another study, Shaffer et al. (2019) analyzed low-frequency variants obtained from SNP arrays and observed a suggestive enrichment for variants in a putative craniofacial enhancer on chromosome 9q22.33.

The present report describes an alternative framework for the delineation of functional rare candidate variants at GWAS loci (Figure S1), based on the resequencing of large case-control study populations. While whole-genome-sequencing represents the most unbiased, systematic method of sequencing, large-scale use of this approach is currently unaffordable. Targeted resequencing approaches have emerged as an alternative option. However, the requirement of an a priori selection of candidate regions represents a considerable challenge, in particular as GWAS loci can encompass several 100 kb. Here, we propose a framework for nsCL/P that first use functional data to identify relevant candidate regions at GWAS loci, including chromatin accessibility and histone modifications in tissues of relevance to craniofacial development, such as human neural crest cells (hNCCs; Rada-Iglesias et al., 2012) and human craniofacial tissue (Wilderman et al., 2018). Subsequently, resequencing is performed using single-molecule molecular inversion probes (smMIPs) in a multiethnic case-control study population. SmMIPs-based resequencing has previously been shown to be cost-efficient, and reliable in the variant calling (Neveling et al., 2017). Both factors are important for the analysis of large-scale study populations when laboratory-based verification of all detected variants is virtually impossible.

We apply this framework to three GWAS candidate loci for nsCL/P, which had shown (i) suggestive evidence of association in prior GWAS and (ii) independent replication in different ethnicities (Ludwig et al., 2017). Identified rare variants were annotated using diverse noncoding variant annotation tools, analyzed for inheritance patterns and combined for burden analyses. Following these different lines of evidence, 16 variants with a putative role in the etiology of nsCL/P were identified.

2 | MATERIALS AND METHODS

2.1 | Study population

The present study was approved by the ethics committees of the respective medical faculties, and all participants provided written informed consent before inclusion. The study population comprised a total of 1061 nsCL/P patients from Bonn, Yemen, and Mexico, and 1591 ethnically matched controls (Table 1). The Bonn case-control sample comprised 694 independent nsCL/P patients of Central European (CEU) ethnicity, and 858 unaffected volunteer blood donors (Mangold et al., 2010). Of the 694 patients, 398 (57%) had been included in a previous GWAS (Mangold et al., 2010). A detailed description of recruitment strategies and DNA extraction in the Bonn sample is provided elsewhere (Mangold et al., 2009). For around 70% of the Bonn sample, DNA was available from first-degree relatives. This includes 432 complete patient-parent trios, allowing for the identification of putative de novo variants (Table S1). A total of 135 patients from the Bonn sample were members of multiply affected families and could be used for co-segregation analyses. Recruitment strategies for the case-control study populations from Mexico (150 nsCL/P patients, 312 controls; Rojas-Martinez et al., 2010) and Yemen (217 nsCL/P patients, 421 controls; Aldhorae et al., 2014) are described in the respective reports. For these two-study populations, no additional family members were available.

2.2 | Selection of candidate regulatory regions (CRRs)

We selected three loci based on recent data from our in-house nsCL/P meta-analysis (Ludwig et al. 2017). Each locus had shown suggestive evidence of association with nsCL/P ($5 \times 10^{-08} < P < 5 \times 10^{-06}$) and, importantly, an association of the common risk variants had been replicated in the same multiethnic sample as used in the present study (Table 2). The latter supports these loci as true nsCL/P risk regions across ethnicities and increases the a priori chance of detecting rare variants at these loci in different populations.

Next, for each of the loci, the 99% credible SNP set was defined based on the respective lead SNP, as described previously (Ludwig et al., 2017; Wellcome Trust Case Control Consortium, 2012). The region was then extended by ±10 kb to allow the inclusion of functional regions at each end. At each locus, functional CRRs were identified based on information on vertebrate sequence conservation and epigenetic data of relevance to craniofacial development. For humans, the following data were accessed: data on chromatin conformation (topologically associated domains [TADs]; human embryonic stem cells (GEO: GSE35156); DNase I hypersensitive sites from embryonic facial prominences (GEO: GSE90336); chromatin states from embryonic craniofacial tissue (GEO: GSE97752); and both histone modifications (H3K4me1, H3K4me3, H3K27ac) and transcription factor binding profiles (EP300, TFAP2A) from hNCCs (GEO: GSE28876, GSE24447). These data were complemented by mouse craniofacial data concerning histone modifications (H3K4me1, H3K4me3, H3K27ac; E11.5 facial processes, E13.5 and E14.5 palatal shelves; Dixon Laboratory, unpublished data); and p300-binding profiles (GEO: GSE49413; Attanasio et al., 2013).

2.3 | Confirmation of enhancer activity for 8q21.11_CRR1

To obtain some insight into whether our approach identifies true functionally relevant elements, one approximately 1 kb large region from chromosome 8q21.11_CRR1 (chr8:76581086-76582089; hg19) was tested in a *lacZ*-based enhancer assay ($n = 1$). We generated a transgene

TABLE 1 Overview of the study cohorts

Ethnicity	nsCL/P patients	Controls	References
Central European	694	858	Mangold et al. (2010)
Mexican	150	312	Rojas-Martinez et al. (2010)
Yemeni	217	421	Aldhorae et al. (2014)

Abbreviation: nsCL/P, nonsyndromic cleft lip with or without cleft palate.

TABLE 2 Overview of loci included in the present study

Locus	Lead SNP GWAS	99% Credible SNP region (hg19)	Adjacent protein-coding genes[a]	Number of functional candidate regions	Total size of sequenced regions	Total number of smMIPs
2q35	rs112800917	220,603,899–220,689,584	*TMEM198, OBSL1, INHA, STK11IP, SLC4A3, EPHA4*	5	14,030 bp	117
8q21.11	rs10808812	77,383,882–77,684,833	*ZFHX4, PEX2*	4	34,765 bp	299
9q22.2	rs4132699	92,170,516–92,291,794	*CKS2, SECISBP2, SEMA4D, GADD45G*	3	7895 bp	80

Abbreviations: GWAS, genome-wide association study; smMIP, single-molecule molecular inversion probes; SNP, single-nucleotide polymorphism.

[a]All genes within the same topologically associated domain, based on data from human embryonic stem cells (GEO: GSE35156).

destination vector based on Poulin et al. (2005), which comprised the LacZ reporter gene with a minimal Hsp68 promoter downstream of AttP flanked ccdB gene for Gateway cloning. The candidate enhancer sequence was amplified with AttB tags and BP Clonase shuttled into the destination vector. After sequence verification, the plasmid was linearized by SphI digestion followed by sepharose column purification in sterile injection buffer (10 mM Tris (pH 7.5), 0.1 mM ethylenediaminetetraacetic acid (pH 8.0), 100 mM NaCl) (Gong and Yang, 2005). DNA was injected into C57BL6/J (Envigo) zygote pronuclei using standard protocols (DeMayo et al., 2012), at 2ng/μl. Zygotes were cultured overnight and the resulting two-cell embryos were surgically implanted into the oviduct of Day 0.5 postcoitum pseudopregnant mice. Embryos were harvested at embryonic day (E)11.5 and E13.5 and stained and genotyped as previously described (Poulin et al., 2005).

2.4 | Design of smMIPs

SmMIP design was performed using an in-house pipeline based on the program "MIPgen" (Boyle et al., 2014), with the following modifications: -min_capture_size 160; -max_capture_size 195; -ext_min_length 13 (average target region per smMIP: 144 bp). SmMIPs comprised a 35 bp backbone (including a 5-bp tag of "N" nucleotides), and variable-length extension and ligation arms. For common variants (dbSNP b150) present in extension or ligation arms, a second smMIP was designed toward the respective variant allele. During quality control (QC), smMIPs with high extension/ligation arm copy numbers (product of both >50), and smMIPs with logistic scores <0.3, were removed and redesigned. Following visual inspection of the smMIPs in the UCSC Genome Browser (Kent et al., 2002), oligonucleotides (100 μM) were ordered from Integrated DNA Technologies (IDT).

2.5 | Library preparation

DNA samples were prepared from 20 ng/μl dilutions, and the presence of double-stranded genomic DNA concentration was verified (Quant-iT PicoGreen dsDNA Assay; Thermo Fisher Scientific). Reaction input was 100 ng. Individual smMIPs were pooled and phosphorylated as described elsewhere (Eijkelenboom et al., 2016), with minor modifications. The smMIP to DNA ratio was set at 800:1. To adjust for over- and under-performing smMIPs, two rebalancing rounds were performed using up to five test samples and the Illumina MiSeq system. The final smMIP pool was then generated and phosphorylated for use in the high-throughput stage of the study.

Library preparation was performed as described elsewhere (Eijkelenboom et al., 2016), with the following modifications: hybridization was performed with 0.03 μl of dNTPs (0.25 mM) and incubated at 60°C for 22.5 h. PCR was performed in a total volume of 25 μl, which contained 5 μl of the exonuclease-treated product; 2x iProof HF Master Mix (Bio-Rad Laboratories); 0.125 μl forward primer (100 μM, IDT); and 1.25 μl sample-specific barcoded reverse primer (10 μM, IDT). Forward and reverse primers specific to

Illumina sequencers were used (O'Roak et al., 2012). The PCR program was as follows: 98°C for 30 s; 20x (98°C for 10 s, 60°C for 30 s, 72°C for 30 s); 74°C for 2 min; and 4°C indefinitely.

2.6 | Sequencing

PCR products were combined in individual pools containing 88 samples each. These were then purified with a 0.7x volume of Agencourt AMPure XP beads (Beckman-Coulter). The yield and the purity of the pools were assessed on an Agilent 2200 TapeStation system using D1000 ScreenTapes. Sets of four pools each were combined into individual *mega-pools* and sequenced on an Illumina HiSeq. 2500 devices in high output mode (2 × 125 bp each), respectively. Before sequencing, standard Illumina sequencing primers were replaced by custom sequencing primers (IDT; O'Roak et al., 2012), in accordance with Illumina's protocol.

2.7 | Data analysis

Base call files from the sequencing devices were demultiplexed and converted into FASTQ files using the bcl2fastq conversion software. The following nonstandard parameters were applied: --no-eamss, --mismatches 1. Variant-calling was performed as described in Hiatt et al. (2013), with minor modifications. Briefly, paired-end reads were merged using Paired-End reAd mergeR (PEAR; Zhang et al. 2014), and aligned to the hg19 reference genome using BWA-MEM (Li and Durbin, 2009). Trimming and collapsing of smMIPs were performed using available MIPgen scripts (Boyle et al., 2014). During this step, reads carrying the same molecular tag (introduced during the synthesis of the smMIPs as five degenerate bases) were collapsed into single reads, thus representing individual hybridization events. This approach creates higher quality consensus sequences and reduces PCR artifacts. Variant calling was performed using GATK UnifiedGenotyper (Van der Auwera et al., 2013). Annotation of the retrieved variants was performed using ANNOVAR (Wang et al. 2010). Further data processing and annotation were performed using samtools 0.1.19 (Li et al., 2009); vcftools 0.1.15 (Danecek et al., 2011); and the tidyverse 1.2.1 package collection in RStudio 1.0.143 (R version 3.4.0).

Samples were excluded when less than 90% of the target region reached at least 30x of collapsed coverage. For variant-level analysis, variants were excluded if they had a quality-by-depth of less than 10 (conservative threshold as suggested by Hiatt et al., 2013, to reduce the number of false positives); or when genotype calls were generated in less than 90% of the samples. In addition, candidate variants were visually inspected in the .bam-files using Integrative Genomics Viewer (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

2.8 | Annotation of noncoding variants

Annotation scores for noncoding variants (Table 3) were downloaded from the respective websites, and integrated into the ANNOVAR variant annotation pipeline (Wang et al., 2010). To assess the degree of redundancy between the individual tools, the correlation between the five different annotation scores was calculated using Spearman's rank correlation coefficient for pairwise complete observations.

TABLE 3 In silico prediction tools for noncoding variants used in the present study

Annotation score	Annotation score reference	Threshold[a]
CADD 1.3	Kircher et al. (2014)	≥15
LINSIGHT[b]	Huang et al. (2017)	≥0.9
FATHMM-MKL noncoding	Shihab et al. (2015)	≥0.9
DANN	Quang et al. (2015)	≥0.9
ReMM[b]	Smedley et al. (2016)	≥0.9

[a] Thresholds used to predict "likely functionally relevant" status, as based on previously published recommendations (Smedley et al. 2016).

[b] In situations where multiple annotation values were provided (e.g., for some insertion and deletions), only the largest value was retained.

2.9 | Combined burden and variance-component test

To test the association between rare variants in each CRR and nsCL/P, the optimal sequence kernel association test SKAT-O (Lee et al., 2012) was used. SKAT-O combines burden and variance component analysis to test for associations that are robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants (Lee et al., 2014). To avoid any bias secondary to population structure, a region-based association test was performed. This involved a separate analysis of the nsCL/P CRRs for each of the three study populations (CEU, Mexican, and Yemeni). Variants used for this were only filtered for QC criteria. The analyses were performed using R package SKAT (version 1.3.2.1) with default parameters, which included default minor allele frequency weighing. The results of the test were reported as p values obtained with the parametric bootstrap option ($n = 1000$). To correct for multiple testing, adjusted p values were calculated by using the Holm-Bonferroni method (Holm, 1979).

2.10 | Selection of putative functional variants in individual families

For the Bonn study population, additional analyses were performed to identify rare variants with a potentially strong functional impact in individual families. Variants that met the following criteria were prioritized: (i) a maximum frequency in controls of 0.1% (selected based on nsCL/P prevalence in the general population); (ii) above-threshold functional impact for at least one score (Table 3); and (iii) validated via Integrative Genomics Viewer visual inspection. To identify de novo and co-segregating variants, samples from available relatives of the affected individuals were analyzed using Sanger sequencing. In addition, to determine the potential of each variant on potential transcription factor binding sites, in silico annotation was performed using RegulomeDB (Boyle et al., 2012).

3 | RESULTS

3.1 | Identification of functional CRRs at three nsCL/P risk loci

Based on the credible SNP set definition (Tables S2 and S3), the sizes of the target regions were 106 kb (2q35); 321 kb (8q21.11); and 121 kb (9q22.2). Integrating functional datasets within each of these regions, we identified several smaller regions with evidence of functional activity. Prioritizing those regions with robust evidence from different datasets revealed 5 (2q35), 4 (8q21.11), and 3 (9q22.2) CRRs (Table 2 and Table S4), which encompass 56.7 kb in total. An exemplar plot for the 8q21.11 locus is provided in Figure 1, while plots for 2q35 and 9q22.2 are provided in Figures S2 and S3, respectively. Initial results from a lacZ-based enhancer assay in embryonic mice indicate that the 1 kb region within 8q21.11_CRR1 is an active enhancer in the nasal epithelium at E11.5. At E13.5, enhancer activity is observed in the midline nasal tissue, in addition to the neural tube and brain structures (Figure S4).

3.2 | Identification and annotation of rare variants within candidate regions

A set of 496 smMIPs was designed to cover the entire set of target sequences (56.7 kb). Upon assay optimization, overall probe performance showed balanced performance and even coverage of all

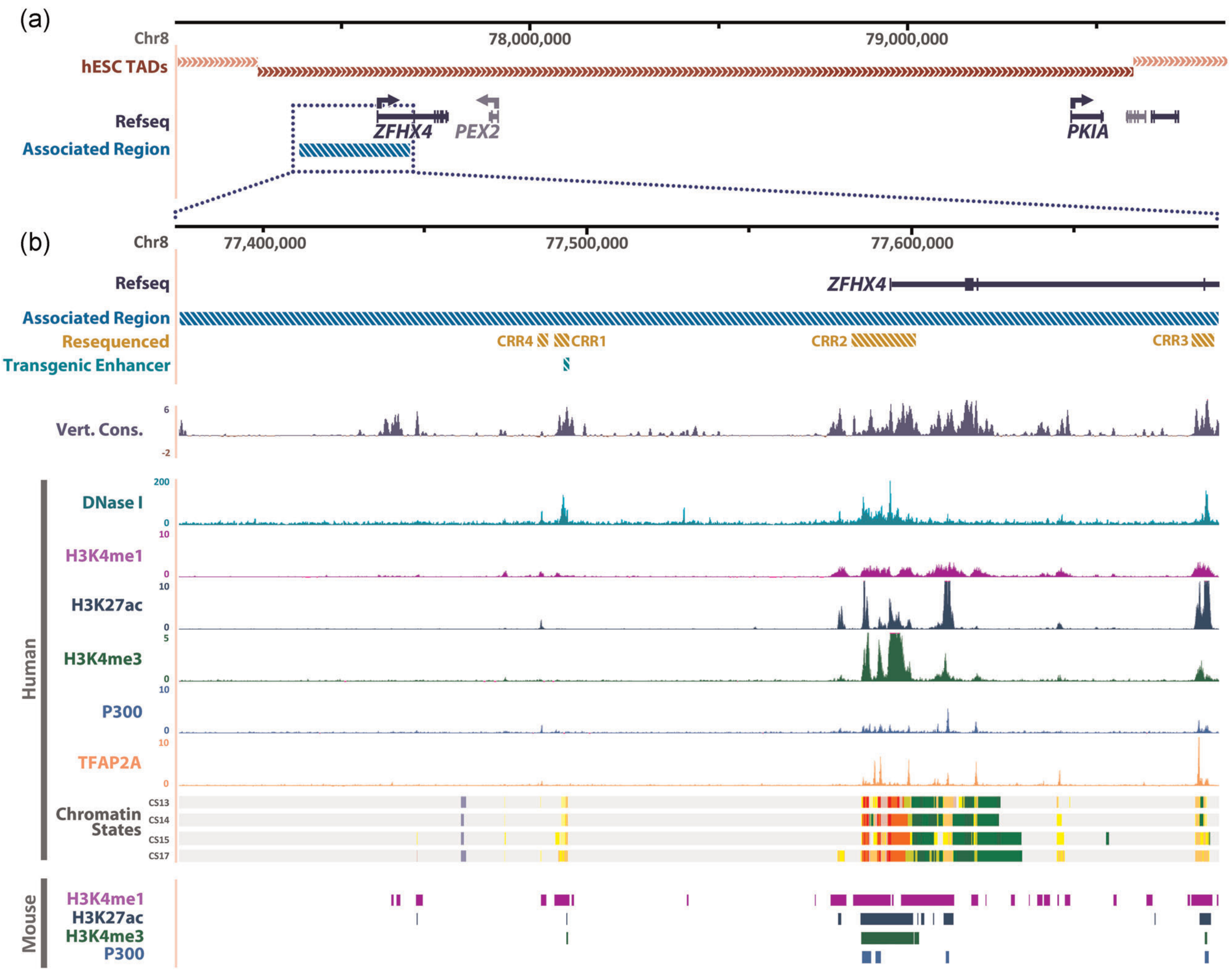

FIGURE 1 Identification of functional candidate regions at the 8q21.11 risk locus. (a) The credible SNP region at 8q21.11 is indicated in blue as an "Associated region," within the context of neighboring genes and the topologically associated domain (TAD) from hESC. (b) Zoom-in of the associated region depicted in (a), showing tracks of the functional data used for the selection of candidate regions. These included (from top to bottom): conservation in vertebrates, DNaseI hypersensitive sites in facial prominences, chromatin modifications in neural crest cells, chromatin states in facial tissue, and data from mouse palatal shelves. Candidate regulatory regions (CRR) for resequencing are highlighted in brown shading (CRR1 through CRR4)

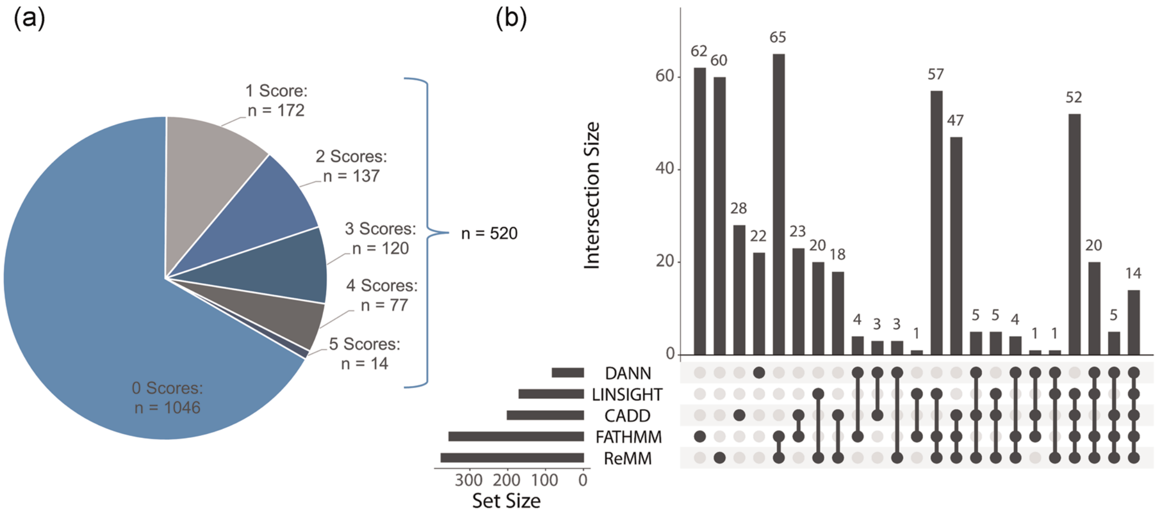

FIGURE 2 Overview of variants exceeding the thresholds set for each of the annotation scores individually and in combinations. (a) number of variants that reach the annotation score threshold in 0–5 annotation scores (independent of which annotation score). (b) UpSet plot illustrating the overlap between the individual annotation scores; overlap categories with no entries are omitted. Set size: number of variants exceeding the set threshold per annotation score. Intersection size: number of variants exceeding the set thresholds for combinations of annotation scores

12 CRRs (Figures S5 and S6). Of the 2652 samples, a total of 2630 fulfilled QC criteria (>99%; 19 controls and 3 patients were excluded). Mean coverage was 1093x before and 242x after collapsing into single-molecule reads. Overall, 2181 variants were identified. Of these, 1566 passed variant-level QC. Upon in silico variant annotation, noncoding annotation scores were correlated across the variant set (Figure S7). None of the pairwise combinations showed a correlation of more than 0.9, the strongest correlation was found between the FATHMM-MKL noncoding and the LINSIGHT scores (0.80). The weakest correlation was found between the FATHMM-MKL noncoding and the DANN scores (0.50). Overall, the majority of variants did not show an above-threshold score for any of the five annotation tools, while 14 variants met this criterion in all five scores (Figure 2 and Table S5).

3.3 | Variant prioritization and pedigree analysis

In the Bonn study population, 964 post-QC variants were identified. Of these, 667 were observed in patients, with a subset of 289 being rare in the in-house controls. The annotation pipeline returned 102 variants (in 116 families) with high scores according to at least one in silico prediction tool (Figure S8). A total of 99 of the 102 variants (97.0%) were validated by Sanger sequencing. Annotation with RegulomeDB identified 12 variants with a score of 2a or 2b, indicative of a likely effect on transcription factor binding (Figure S9 and Table S6). Eight of these 12 variants mapped to CRR2 at 8q21.11, including four variants that were located within a 1 kb interval. Among those is one variant (NC_000008.10:g.77585212G>A) that is predicted to disrupt the binding motif of Pitx2, a transcription factor with an established role in murine palatogenesis (Lu et al., 1999).

Of the 79 variants that underwent pedigree analysis, the majority (*n* = 75) were present in at least one unaffected family member. This also includes 11 of the 12 variants with high RegulomeDB scores. Four variants were carried by at least one additional affected family member, were not present in additional unaffected family members and absent from gnomAD (three variants) or reported with a frequency below 0.01% (one variant, Table S6). These four variants mapped to two of the three GWAS loci (2q35, 8q21.11, Table 4), in different CRRs.

TABLE 4 Results of the combined burden and variance-component analyses

Region	*p* Value (Central European)	*p* Value (Mexican)	*p* Value (Yemeni)
2q35_CRR1	.836	.160	.539
2q35_CRR2	**.028**	.117	.130
2q35_CRR3	.437	**.015**	.330
2q35_CRR4	.119	.067	.515
2q35_CRR5	.366	.183	.618
8q21.11_CRR1	.056	**.049**	.441
8q21.11_CRR2	.262	.158	.697
8q21.11_CRR3	.924	**.007**	.112
8q21.11_CRR4	.124	1	.562
9q22.2_CRR1	**.039**	.922	.714
9q22.2_CRR2	.781	.840	.227
9q22.2_CRR3	1	.721	.682

Note: Nominally significant results are indicated in bold.

3.4 | Combined burden and variance-component test

For each CRR and each study population, a combined burden and variance component analysis was performed. No CRR was significantly enriched for rare variants after correction for multiple testing (Table 4). Nominal significance ($p < .05$) was observed for regions 2q35_CRR2 and 9q22.2_CRR1 in the Bonn study population, and 2q35_CRR3, 8q21.11_CRR1, and 8q21.11_CRR3 in the Mexican study population.

4 | DISCUSSION

The present report suggests a framework for the parallel identification of private and rare variants that might contribute to the etiology of nsCL/P. This approach involves an a priori definition of CRRs, their cost-efficient resequencing in large case-control study populations, and subsequent pedigree- and burden analyses. Although multiple exome- and candidate-gene sequencing studies of nsCL/P have been recently performed to shed light on the contribution of rare variants in protein-coding genes, few investigations have focused on those located in noncoding regions. Leslie et al. (2015) sequenced targeted GWAS regions in 1498 nsCL/P patient-parent trios and identified multiple de novo variants that represent candidate variants for nsCL/P. In that study, the lack of functional data integration resulted in the sequencing of a large fraction without a functional role. Using a study population of 2216 patients with OFC and 1576 controls, Shaffer et al. (2019) genotyped around 16,000 low-frequency variants in regions of functional relevance to craniofacial development (i.e., putative craniofacial enhancers). Although the authors demonstrated suggestive enrichment for one element on 9q22.33, no follow-up pedigree analyses were performed to identify rare variants with large effect sizes in individual families. The design of this latter study precluded the analysis of private variants, and the results may have been biased by the limited quality of low-frequency variants obtained from SNP arrays (Wright et al., 2019). In our present approach, we combine the advantages of each of the two studies. First, CRRs were selected for resequencing using diverse annotations from publicly available datasets that are of relevance to craniofacial development. The results of the in vivo assay supported our in silico strategy, even though we cannot entirely rule out a potential positional effect due to a limited number of embryos analyzed. While we expect this approach to increase the chance of identifying functionally relevant variants cost-efficiently, some variants might have been missed due to the current incompleteness of regulatory maps of craniofacial tissues and the manual assignment of CRRs. The former limitation might be addressed by increasing the availability of datasets in other tissues/cell systems. The latter warrants a bioinformatic solution with sensitive parameters to balance between functionally relevant and nonrelevant annotations. Other systematic approaches, such as integrating RegulomeDB scores, have also been suggested (Jones et al., 2019). In the present study, we investigated three GWAS loci at which common variation was found to be associated across different populations (Ludwig et al., 2017). However, we anticipate that our strategy can also be applied to populations that have not (yet) shown robust associations in GWAS, as rare variants do not share strong LD with common variants and can, therefore, exert their effects in a biologically independent manner.

Technically, our resequencing data confirmed the previously demonstrated robustness of the smMIP method (Eijkelenboom et al., 2016), indicated by a low number of samples failing QC and a mean coverage that was comparable across all regions. For variant calling, we had set the QC at conservative levels to minimize the risk of false positives. The high validation rate observed in our study, which corresponds to a previously reported positive predictive value of 98% (O'Roak et al., 2012), confirms this strategy and allowed for the identification of a large number of variants with sufficient confidence. After QC, we identified over 1500 variants across the entire dataset, with a linear correlation found between the number of variants and the size of the target region. As in most rare variant studies, the inherent challenge is to distinguish between benign variants and rare variants that contribute to nsCL/P risk. While protein-coding variants can be annotated based on their effect on protein sequence and/or structure, no systematic "regulatory code" has yet been established for noncoding variants. Moreover, since regulatory elements are often tissue-specific, functional effects can be assumed to vary between different tissues, which increases the demand for tissue-specific annotation scores. Indeed, previous authors have suggested that no current annotation score for noncoding regions performs well in all circumstances (Kircher et al., 2019), and in silico experiments have shown that this limitation reduces the power to identify noncoding regulatory elements of relevance to specific diseases (Short et al., 2018). Our correlation analyses—which confirmed previous observations—and other available data suggest that variant prioritization should include complementary approaches, for which the presence of a single high annotation score would be sufficient to prioritize a variant. However, this also means an increase in the number of potentially causative variants, and a substantial fraction of these would represent variants likely not associated with the disease under study. While we tried to reduce the number of those "false positives" using current state-of-the-art tools (e.g., annotation/frequencies), our framework is scalable to include additional measures when they become available, which would then allow for better discrimination between benign and deleterious variants.

In the present study, we used two approaches to narrow down potentially relevant variants. First, a family-based analysis was performed, grounded on the hypothesis that high-penetrance private variants will be present in individual families, as shown recently for noncoding regions around *IRF6* (Fakhouri et al., 2014), *FZD6* (Cvjetkovic et al., 2015), and *FGFR2* (Leslie et al., 2015). We did not detect any de novo variant in our study, which may be attributable to two reasons. First, the limited size of the sequenced regions, which might have missed some etiological variants that are located outside

TABLE 5 Variants with high RegulomeDB scores and co-segregating variants

Chromosome	Position (hg19)	Reference allele	Alternative allele	HGVS nomenclaturea	RegulomeDB score	Family analysis
2	220,611,617	A	C	NC_000002.11:g.220611617A>C	5	Co-segregating
2	220,637,204	C	T	NC_000002.11:g.220637204C>T	3a	Co-segregating
8	77,486,006	C	T	NC_000008.10:g.77486006C>T	5	Co-segregating
8	77,491,677	CTAAA	–	NC_000008.10:g.77491684_77491688delAAACT	5	Co-segregating
2	220,636,955	C	T	NC_000002.11:g.220636955C>T	2b	Reduced penetrance
8	77,492,641	C	G	NC_000008.10:g.77492620G>A	2b	Reduced penetrance
8	77,493,628	–	A	NC_000008.10:g.77493635_77493635dupA	2b	Reduced penetrance
8	77,585,212	G	A	NC_000008.10:g.77585212G>A	2b	NA
8	77,585,220	A	C	NC_000008.10:g.77585220A>C	2b	Reduced penetrance
8	77,585,282	G	T	NC_000008.10:g.77585282G>T	2b	Reduced penetrance
8	77,586,085	G	A	NC_000008.10:g.77586089delG	2a	Reduced penetrance
8	77,594,281	C	T	NC_000008.10:g.77593754A>G	2b	Reduced penetrance
8	77,595,781	C	T	NC_000008.10:g.77595781C>T	2b	Reduced penetrance
8	77,595,782	–	C	NC_000008.10:g.77595789_77595789dupC	2b	Reduced penetrance
8	77,599,296	A	C	NC_000008.10:g.77599296A>C	2b	Reduced penetrance
8	77,689,172	G	C	NC_000008.10:g.77689172G>C	2b	Reduced penetrance

of the credible SNP set region or the same TAD as the GWAS locus. Second, parental DNA was only available for a limited number of variants, thus potential de novo variants might still be present among the variants but have not been analyzed (despite the availability of DNA from both parents for 62.5% of the cohort). As expected, the vast majority of variants were also present in unaffected family members. This either suggests reduced penetrance for this variant, which might be driven by the polygenic background or a second hit in the coding region on the same haplotype (Castel et al., 2018). Alternatively, these variants do not contribute to nsCL/P. We also detected six variants that were not present in seven additional affected family members but observed an equal or larger number of affected structures in six out of the seven comparisons. Notably, we also identified four variants that occurred in all affected relatives (and none of the unaffected relatives) within the respective families, and an additional set of 12 variants that showed reduced (n = 11) or inconclusive (n = 1) penetrance but strong functional support from in silico scores and RegulomeDB. The latter included a variant that is predicted to disrupt the binding motif of Pitx2, a transcription factor with an established role in murine palatogenesis (Lu et al., 1999). While the number of co-segregating variants is in line with the number expected by chance, overall, these 16 variants represent candidates for functional studies, which are beyond the scope of this study (Table 5). While these analyses can help to decipher the pathomechanism at the respective loci, the translation of these findings to individual counseling remains limited.

The second approach involved the investigation of enrichment of rare variants within individual CRRs, to highlight potential relevant functional elements. No test-wide enrichment of rare variants was found. In three regions, this analysis revealed nominally significant findings, which is consistent with previous findings from Shaffer et al. (2019), who also generated nominally significant results only, and emphasizes the need for the investigation of larger sample sizes. Notably, in all three regions, rare variants were more frequent among controls, suggesting that rare variants are protective in these regions, but this hypothesis warrants further investigation. Importantly, functional regions with an excess of rare variants that obtain further support by independent studies represent good candidates for future integration into diagnostic gene panels, provided that improved annotation and interpretation scores for noncoding variants become available.

In summary, the present study demonstrated the feasibility of a novel framework for the further analysis of risk loci for multifactorial diseases. By combining GWAS results with functionally relevant data, candidate CRRs were selected, and resequenced in a large multiethnic study population to identify rare, putative functionally relevant variants. In its current form, our framework is designed to detect rare variants at noncoding GWAS loci and could be applied to other traits, in particular developmental diseases in which (i) access to relevant tissue is limited; and (ii) a contribution of rare variants is hypothesized. However, the approach could also be extended to other sets of noncoding CRRs, such as craniofacial-specific enhancers (Wilderman et al., 2018), and to include the TADs of all genome-wide significant and suggestively significant risk loci. In our study, we identified 16 rare variants with high in silico annotation scores, which now represent promising candidates for future functional analyses, such as quantification of the allele-specific effects on regulatory activity and the identification of the biological effect on downstream genes. While it is anticipated that whole-genome sequencing (WGS) will make a preselection of CRRs and the resequencing strategy obsolete in the future, our present strategy is highly valuable until computational and financial constraints that currently prevent WGS to be performed in large cohorts are overcome.

ACKNOWLEDGMENTS

The authors thank all patients, relatives, and control individuals for their participation. We thank Prof. Bernd Pötzsch for assistance in the collection of DNA samples from anonymous blood donors. We acknowledge the contribution of our collaborating clinical partners, laboratory technicians, and colleagues with responsibility for database management. We thank the German support group for individuals with cleft lip and/or palate (Wolfgang Rosenthal Gesellschaft) for assistance with recruitment, and the West German Genome Center for access to sequencing facilities. We thank Dr. Neil Humphreys and the Genome Editing unit for technical support. We thank Martin Kircher for the helpful discussions. The authors have no potential conflicts of interest to declare with respect to the authorship and/or publication of this article. This work was supported by the German Research Foundation (DFG, Grant LU 1944-3/1 to KUL) and the European Science Foundation. Open Access funding enabled and organized by Projekt DEAL.

WEB RESOURCES

UCSU Genome Browser: https://genome-euro.ucsc.edu/
ANNOVAR: https://doc-openbio.readthedocs.io/projects/annovar/en/latest/
CADD: https://cadd.gs.washington.edu/
DANN: https://cbcl.ics.uci.edu/public_data/DANN/
ReMM: https://charite.github.io/software-remm-score.html
LINSIGHT: http://compgen.cshl.edu/LINSIGHT/
FATHMM-MKL: http://fathmm.biocompute.org.uk/downloads.html
gnomAD: https://gnomad.broadinstitute.org/
MIPgen: https://github.com/shendurelab/MIPGEN/
Mutalyzer: https://mutalyzer.nl/
Samtools: http://www.htslib.org/
BWA-MEM: https://github.com/lh3/bwa
SKAT: https://cran.r-project.org/web/packages/SKAT/index.html
RegulomeDB: https://www.regulomedb.org/regulome-search/

DATA AVAILABILITY STATEMENT

Summary variant information is provided as Supplement (genotype distribution per cohort; Tables S7–S9). Information on the 16 variants of interest is available at https://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1063149.

ORCID

Frederic Thieme http://orcid.org/0000-0001-9839-1764
Kerstin U. Ludwig http://orcid.org/0000-0002-8541-2519

REFERENCES

Aldhorae, K. A., Böhmer, A. C., Ludwig, K. U., Esmail, A. H., Al-Hebshi, N. N., Lippke, B., Gölz, L., Nöthen, M. M., Daratsianos, N., Knapp, M., Jäger, A., & Mangold, E. (2014). Nonsyndromic Cleft Lip with or without Cleft Palate in Arab populations: Genetic analysis of 15 risk loci in a novel case–control sample recruited in Yemen. *Birth defects research. Part A, Clinical and molecular teratology*, *100*(4), 307–313.

Attanasio, C., Nord, A. S., Zhu, Y., Blow, M. J., Li, Z., Liberton, D. K., Morrison, H., Plajzer-Frick, I., Holt, A., Hosseini, R., Phouanenavong, S., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., FitzPatrick, D. R., Ren, B., Hallgrímsson, B., Pennacchio, L. A., & Visel, A. (2013). Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, *342*(6157), 1241006.

Basha, M., Demeer, B., Revencu, N., Helaers, R., Theys, S., Bou Saba, S., Boute, O., Devauchelle, B., Francois, G., Bayet, B., & Vikkula, M. (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. *Journal of Medical Genetics*, *55*, 449–458.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., & Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797.

Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A., & Shendure, J. (2014). MIPgen: Optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics*, *30*(18), 2670–2672.

Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., & Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, *50*(9), 1327–1334.

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., Yang, Y., Carstens, R. P., Anand, D., Lachke, S. A., Smith, J. D., Dorschner, M. O., Bedell, B., Kirk, E., Hing, A. V., … Roscioli, T. (2018). Mutations in the epithelial cadherin-p120-catenin complex cause mendelian non-syndromic Cleft Lip with or without Cleft palate. *American Journal of Human Genetics*, *102*(6), 1143–1157.

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., & Hecht, J. T. (2015). Regulatory variant in *FZD6* gene contributes to nonsyndromic cleft lip and palate in an African-American family. *Molecular Genetics & Genomic Medicine*, *3*(5), 440–451.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R., 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158.

DeMayo, J. L., Wang, J., Liang, D., Zhang, R., & Demayo, F. J. (2012). Genetically engineered mice by pronuclear DNA microinjection. *Curr Protoc Mouse Biol*, *2*(3), 245–262.

Eijkelenboom, A., Kamping, E. J., Kastner-van Raaij, A. W., Hendriks-Cornelissen, S. J., Neveling, K., Kuiper, R. P., Hoischen, A., Nelen, M. R., Ligtenberg, M. J., & Tops, B. B. (2016). Reliable next-generation sequencing of formalin-fixed, paraffin-embedded tissue using single molecule tags. *Journal of Molecular Diagnostics*, *18*(6), 851–863.

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., Leslie, E., Pennacchio, L. A., Van Bokhoven, H., Visel, A., Zhou, H., Murray, J. C., & Schutte, B. C. (2014). An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. *Human Molecular Genetics*, *23*(10), 2711–2720.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., GTEx, C., Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., McCarthy, M. I., Dermitzakis, E. T., Cox, N. J., & Ardlie, K. G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics*, *50*(7), 956–967.

Gong, S., & Yang, X. W. (2005). Modification of bacterial artificial chromosomes (BACs) and preparation of intact BAC DNA for generation of transgenic mice. *Curr Protoc Mouse Biol*, *31*(1), 5.21.21–25.21.13.

Grosen, D., Bille, C., Petersen, I., Skytthe, A., Hjelmborg, J., Pedersen, J. K., Murray, J. C., & Christensen, K. (2011). Risk of oral clefts in twins. *Epidemiology*, *22*(3), 313–319.

Hiatt, J. B., Pritchard, C. C., Salipante, S. J., O'Roak, B. J., & Shendure, J. (2013). Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Research*, *23*(5), 843–854.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Huang, Y. F., Gulko, B., & Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, *49*(4), 618–624.

Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, *48*(2), 214–220.

Jones, S. A., Cantsilieris, S., Fan, H., Cheng, Q., Russ, B. E., Tucker, E. J., Harris, J., Rudloff, I., Nold, M., Northcott, M., Dankers, W., Toh, A., White, S. J., & Morand, E. F. (2019). Rare variants in non-coding regulatory regions of the genome that affect gene expression in systemic lupus erythematosus. *Scientific Reports*, *9*(1), 15433.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*, 996–1006.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315.

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J., & Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, *10*(1), 3583.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, *95*(1), 5–23.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X., NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, *91*(2), 224–237.

Leslie, E. J., Carlson, J. C., Shaffer, J. R., Butali, A., Buxó, C. J., Castilla, E. E., Christensen, K., Deleyiannis, F. W., Leigh Field, L., Hecht, J. T., Moreno, L., Orioli, I. M., Padilla, C., Vieira, A. R., Wehby, G. L., Feingold, E., Weinberg, S. M., Murray, J. C., Beaty, T. H., &

Marazita, M. L. (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between *FOXE1* and all orofacial clefts, and *TP63* and cleft lip with or without cleft palate. *Human Genetics*, *136*(3), 275–286.

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., Fulton, R. S., Kousa, Y. A., Fakhouri, W. D., Naji, A., Ruczinski, I., Begum, F., Parker, M. M., Busch, T., Standley, J., … Murray, J. C. (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. *American Journal of Human Genetics*, *96*(3), 397–411.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R., Genome Project Data Processing Subgroup. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Liu, H., Leslie, E. J., Carlson, J. C., Beaty, T. H., Marazita, M. L., Lidral, A. C., & Cornell, R. A. (2017). Identification of common non-coding variants at 1p22 that are functional for non-syndromic orofacial clefting. *Nature Communications*, *8*, 14759.

Lu, M.-F., Pressman, C., Dyer, R., Johnson, R. L., & Martin, J. F. (1999). Function of Rieger syndrome gene in left-right asymmetry and craniofacial development. *Nature*, *401*, 276–278.

Ludwig, K. U., Böhmer, A. C., Bowes, J., Nikolic, M., Ishorst, N., Wyatt, N., Hammond, N. L., Gölz, L., Thieme, F., Barth, S., Schuenke, H., Klamt, J., Spielmann, M., Aldhorae, K., Rojas-Martinez, A., Nöthen, M. M., Rada-Iglesias, A., Dixon, M. J., Knapp, M., & Mangold, E. (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. *Human Molecular Genetics*, *26*(4), 829–842.

Maller, J. B., Mcvean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hall, A. S., … Donnelly, P., Wellcome Trust Case Control Consortium. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N. A., Chawa, T. A., Mattheisen, M., Steffens, M., Barth, S., Kluck, N., Paul, A., Becker, J., Lauster, C., Schmidt, G., Braumann, B., Scheer, M., … Nöthen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, *42*(1), 24–26.

Mangold, E., Ludwig, K. U., & Nöthen, M. M. (2011). Breakthroughs in the genetics of orofacial clefting. *Trends in Molecular Medicine*, *17*(12), 725–733.

Mangold, E., Reutter, H., Birnbaum, S., Walier, M., Mattheisen, M., Henschke, H., Lauster, C., Schmidt, G., Schiefke, F., Reich, R. H., Scheer, M., Hemprich, A., Martini, M., Braumann, B., Krimmel, M., Opitz, C., Lenz, J. H., Kramer, F. J., Wienker, T. F., … Diaz Lacava, A. (2009). Genome-wide linkage scan of nonsyndromic orofacial clefting in 91 families of central European origin. *American Journal of Medical Genetics. Part A*, *149A*(12), 2680–2694.

Marini, N. J., Asrani, K., Yang, W., Rine, J., & Shaw, G. M. (2019). Accumulation of rare coding variants in genes implicated in risk of human cleft lip with or without cleft palate. *American Journal of Medical Genetics. Part A*, *179*, 1260–1269.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195.

Mossey, P. A., & Modell, B. (2012). Epidemiology of oral Clefts 2012: An international perspective. *Frontiers of Oral Biology*, *16*, 1–18.

Neveling, K., Mensenkamp, A. R., Derks, R., Kwint, M., Ouchene, H., Steehouwer, M., van Lier, B., Bosgoed, E., Rikken, A., Tychon, M., Zafeiropoulou, D., Castelein, S., Hehir-Kwa, J., Tjwan Thung, D., Hofste, T., Lelieveld, S. H., Bertens, S. M., Adan, I. B., Eijkelenboom, A., … Hoischen, A. (2017). BRCA testing by single-molecule molecular inversion probes. *Clinical Chemist*, *63*(2), 503–512.

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J. B., Turner, E. H., Levy, R., O'Day, D. R., Krumm, N., Coe, B. P., Martin, B. K., Borenstein, E., … Shendure, J. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorder. *Science*, *338*(6114), 1619–1622.

Poulin, F., Nobrega, M. A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E. M., & Pennacchio, L. A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, *85*(6), 774–781.

Quang, D., Chen, Y., & Xie, X. (2015). DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761–763.

Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S. A., Swigut, T., & Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell*, *11*(5), 633–648.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C., Melbye, M., Jugessur, A., Lie, R. T., Wilcox, A. J., Fitzpatrick, D. R., Green, E. D., Mossey, P. A., Little, J., Steegers-Theunissen, R. P., … Murray, J. C. (2008). Disruption of an AP-2α binding site in an *IRF6* enhancer is associated with cleft lip. *Nature Genetics*, *40*(11), 1341–1347.

Richardson, T. G., Campbell, C., Timpson, N. J., & Gaunt, T. R. (2016). Incorporating non-coding annotations into rare variant analysis. *PLoS One*, *11*(4), e0154181.

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J. M., Kuruvilla, F., … Daly, M. J. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, *43*(11), 1066–1073.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26.

Rojas-Martinez, A., Reutter, H., Chacon-Camacho, O., Leon-Cachon, R. B., Munoz-Jimenez, S. G., Nowak, S., Becker, J., Herberz, R., Ludwig, K. U., Paredes-Zenteno, M., Arizpe-Cantú, A., Raeder, S., Herms, S., Ortiz-Lopez, R., Knapp, M., Hoffmann, P., Nöthen, M. M., & Mangold, E. (2010). Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, *88*(7), 535–537.

Savastano, C. P., Brito, L. A., Faria, A. C., Seto-Salvia, N., Peskett, E., Musso, C. M., Alvizi, L., Ezquina, S. A., James, C., Gosgene, et al. 2016. Impact of rare variants in ARHGAP29 to the etiology of oral clefts: role of loss-of-function vs missense variants. Clin Genet.

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., Orioli, I. M., Padilla, C., Vieira, A. R., Wehby, G. L., Murray, J. C., Weinberg, S. M., Marazita, M. L., & Leslie, E. J. (2019). Association of low-frequency genetic variants in regulatory regions with nonsyndromic orofacial clefts. *American Journal of Medical Genetics. Part A*, *179*(3), 467–474.

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., Gaunt, T. R., & Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, *31*(10), 1536–1543.

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., & Hurles, M. E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, *555*(7698), 611–616.

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N. L., McMurry, J. A., Haendel, M. A., Mungall, C. J., Lewis, S. E., Groza, T., Valentini, G., & Robinson, P. N. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *American Journal of Human Genetics*, *99*(3), 595–606.

Thieme, F., & Ludwig, K. U. (2017). The role of noncoding genetic variation in isolated Orofacial Clefts. *Journal of Dental Research*, *96*(11), 1238–1247.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 1–33.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164.

Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J. P., & Cotney, J. (2018). High-resolution epigenomic atlas of human embryonic craniofacial development. *Cell Reports*, *23*(5), 1581–1597.

Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., Beaumont, R. N., Tyrrell, J., Wood, A. R., Frayling, T. M., Hattersley, A. T., & Weedon, M. N. (2019). Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *American Journal of Human Genetics*, *104*(2), 275–286.

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., … Bian, Z. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nature Communications*, *8*, 14364.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Thieme, F., Henschel, L., Hammond, N. L., Ishorst, N., Hausen, J., Adamson, A. D., Biedermann, A., Bowes, J., Zieger, H. K., Maj, C., Kruse, T., Buness, A., Hoischen, A., Gilissen, C., Kreusch, T., Jäger, A., Gölz, L., Braumann, B., Aldhorae, K., … Ludwig, K. U. (2021). Extending the allelic spectrum at noncoding risk loci of orofacial clefting. *Human Mutation*, 42, 1066–1078. https://doi.org/10.1002/humu.24219