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

Elizabeth J. Leslie,[1,22,*] Margaret A. Taub,[2,22] Huan Liu,[3,4] Karyn Meltz Steinberg,[5] Daniel C. Koboldt,[5] Qunyuan Zhang,[6] Jenna C. Carlson,[7] Jacqueline B. Hetmanski,[8] Hang Wang,[8] David E. Larson,[5] Robert S. Fulton,[5] Youssef A. Kousa,[9] Walid D. Fakhouri,[10] Ali Naji,[10] Ingo Ruczinski,[2] Ferdouse Begum,[8] Margaret M. Parker,[8] Tamara Busch,[11] Jennifer Standley,[11] Jennifer Rigdon,[11] Jacqueline T. Hecht,[12] Alan F. Scott,[13] George L. Wehby,[14] Kaare Christensen,[15] Andrew E. Czeizel,[16] Frederic W.-B. Deleyiannis,[17] Brian C. Schutte,[18] Richard K. Wilson,[5] Robert A. Cornell,[19] Andrew C. Lidral,[3] George M. Weinstock,[5,20] Terri H. Beaty,[8] Mary L. Marazita,[1,21,*] and Jeffrey C. Murray[11]

Although genome-wide association studies (GWASs) for nonsyndromic orofacial clefts have identified multiple strongly associated regions, the causal variants are unknown. To address this, we selected 13 regions from GWASs and other studies, performed targeted sequencing in 1,409 Asian and European trios, and carried out a series of statistical and functional analyses. Within a cluster of strongly associated common variants near *NOG*, we found that one, rs227727, disrupts enhancer activity. We furthermore identified significant clusters of non-coding rare variants near *NTN1* and *NOG* and found several rare coding variants likely to affect protein function, including four nonsense variants in *ARHGAP29*. We confirmed 48 de novo mutations and, based on best biological evidence available, chose two of these for functional assays. One mutation in *PAX7* disrupted the DNA binding of the encoded transcription factor in an in vitro assay. The second, a non-coding mutation, disrupted the activity of a neural crest enhancer downstream of *FGFR2* both in vitro and in vivo. This targeted sequencing study provides strong functional evidence implicating several specific variants as primary contributory risk alleles for nonsyndromic clefting in humans.

Introduction

Genome-wide association studies (GWASs) have collectively identified thousands of genetic risk factors for various complex human diseases. Although the associated SNPs identified through GWASs might themselves be functional, it is likely that many are in linkage disequilibrium with causal variants. It remains a major challenge to identify such causal variants because the most significant SNPs are frequently located in non-coding regions of the genome. A second limitation is that GWASs focus on common variants, even though the genetic architecture underlying complex human traits probably includes a combination of common, rare, and de novo risk alleles. Targeted sequencing of large genomic regions achieving genome-wide significance creates the opportunity to address both of these issues. Requisite for these studies is a detailed catalog of genetic variation at each locus.

We undertook a targeted sequencing study of regions associated with orofacial clefts (MIM 119530), specifically focusing on nonsyndromic cleft lip with or without cleft palate (NSCL/P). NSCL/P affects approximately 1 in 700 live births and exhibits a complex etiology due to multiple genetic and environmental risk factors.[1] As is true for many complex traits, substantial progress in gene identification for NSCL/P has recently occurred, largely as a result

[1]Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA 15219, USA; [2]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA; [3]Department of Orthodontics, College of Dentistry, University of Iowa, Iowa City, IA 52242, USA; [4]State Key Laboratory Breeding Base of Basic Science of Stomatology (Hubei-MOST) and Key Laboratory for Oral Biomedicine of Ministry of Education, School and Hospital of Stomatology, Wuhan University, 430072 Wuhan, China; [5]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA; [6]Department of Statistical Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA; [7]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; [8]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA; [9]Department of Biochemistry and Molecular Biology, College of Osteopathic Medicine, Michigan State University, East Lansing, MI 48824, USA; [10]Department of Diagnostic and Biomedical Sciences, School of Dentistry, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [11]Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA; [12]Department of Pediatrics, University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [13]Institute of Genetic Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA; [14]Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52242, USA; [15]Department of Epidemiology, Institute of Public Health, University of Southern Denmark, 5230 Odense, Denmark; [16]Foundation for the Community Control of Hereditary Diseases, Budapest 1148, Hungary; [17]Department of Surgery, Plastic and Reconstructive Surgery, University of Colorado School of Medicine, Denver, CO 80045, USA; [18]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA; [19]Department of Anatomy and Cell Biology, Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA; [20]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06117, USA; [21]Department of Human Genetics, Graduate School of Public Health, and Clinical and Translational Science Institute, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA
[22]These authors contributed equally to this work
*Correspondence: ejl40@pitt.edu (E.J.L.), marazita@pitt.edu (M.L.M.)
http://dx.doi.org/10.1016/j.ajhg.2015.01.004.

of GWASs.[2] Prior to GWASs only one gene, *IRF6* (MIM 607119), had been consistently shown to have common variants associated with increased risk of NSCL/P.[3,4] After genome-wide linkage studies,[5] four GWASs,[6–9] and a GWAS meta-analysis,[10] 12 loci associated with NSCL/P are now identified and replicated on chromosomes 1p22, 1p36, 2p21, 3p11.1, 8q21.3, 8q24, 9q22, 10q25, 15q22, 17p13, 17q22, and 20q12.

To deepen our understanding of the genetic architecture of NSCL/P, we sequenced more than 4,000 participants for 13 genomic intervals implicated in NSCL/P. We used a case-parent trio design, which has many advantages including improved quality of genotype calls and haplotypes, accurate calling of de novo mutations, and robustness to population stratification in statistical analyses. These 13 regions included 9 regions selected as high-priority candidates from GWASs and/or genome-wide linkage studies and 4 regions containing candidate genes with prior compelling evidence of rare variants contributing to NSCL/P. This is the first study to perform targeted sequencing of NSCL/P GWAS regions and is among the first to sequence the complete GWAS intervals, including non-coding and coding DNA, for any complex disease or trait. Our international cohort of participants consists of case-parent trios, which allows us to accurately identify de novo mutations and makes it possible to search for the contributions of both rare and common variants as risk alleles for NSCL/P. Here we present the highlights of our study that demonstrate the strength of targeted sequencing to identify functional variants.

Subjects and Methods

Samples

The study population included 1,498 case-parent trios recruited by several research groups with samples coming from individuals of Asian or European ancestry from Europe, the United States, China, and the Philippines (Table S1). After QC as described below, 1,409 of the trios (4,227 individuals) were included in the final analyses. Approval for all research work was obtained from the Institutional Review Boards of participating institutions (both US and foreign) and informed consent was obtained from parents of minor children and from all affected individuals old enough to give their own consent. Affected subjects were diagnosed as having cleft lip (CL) or cleft lip with cleft palate (CLP) based on physical exam. Individuals with other congenital anomalies, recognized CL/P syndromes, or developmental delays were excluded from this study.

The 1,409 case-parent trios are from a subset of populations previously studied by a GWAS, although not all sequenced trios were part of the prior GWAS. Previous work on the same populations included principal-component analysis (PCA) to explore genetic distances. In these studies, the Europeans and European Americans formed one cluster and the Asian individuals formed another. Although a separate PCA of founders from the Asian trios (i.e., the parents) appeared to separate the Philippines from other Asian sites, the F_{ST} values were small ($F_{ST} < 0.022$). Given these data, we stratified our cohort into two populations for analyses: an Asian group combining trios from the Philippines and China (1,034 total trios) and a European group combining all European and European American trios (375 total trios).

Selection of Targeted Regions

Thirteen high-priority regions were selected for sequencing (Table 1), representing 6.3 Mb. Nine regions were identified by GWASs[6,9] and/or genome-wide linkage studies[5,11] as primary or secondary genome-wide significant "hits." Four other regions were selected from candidate gene studies where there was evidence for some contribution of rare variants that could be best clarified by sequencing. The coordinates for sequencing in each of these regions were based on the location of the original GWAS SNPs with p values $< 10^{-5}$, LD structure in CEU or CHB/JPT HapMap samples, annotations of regulatory regions from the ENCODE project, and the location of candidate genes in the immediate region. We targeted both coding and non-coding sequence at each locus to select the boundaries of the targeted regions, which totaled 6.3 Mb of sequence.

Sequencing

Illumina multiplexed libraries were constructed with 1 μg of native genomic DNA according to the manufacturer's protocol (Illumina) with the following modifications. (1) DNA was fragmented with a Covaris E220 DNA Sonicator (Covaris) to range in size between 100 and 400 bp. (2) Illumina adaptor-ligated library fragments were amplified in four 50 μl PCR reactions for 18 cycles. (3) Solid phase reversible immobilization (SPRI) bead cleanup was used for enzymatic purification throughout the library process, as well as final library size selection targeting 300–500 bp fragments. We designed NimbleGen (Roche NimbleGen) custom target probes to the 6.6 Mb target region and performed hybrid capture on pools of 96 indexed samples per capture. We then sequenced each capture pool on two lanes of Illumina HiSeq per manufacturer's recommendations (Illumina) for an average of ~40 Gb per lane or ~835 Mb per sample.

Reads were mapped to the GRCh37-lite reference sequence by BWA v.0.5.9[12] with the following parameters: −t 4 −q 4. Alignments were merged and duplicates marked by Picard v.1.46. Germline and de novo variant calling was performed with Polymutt (v.0.11). Polymutt employs a likelihood-based framework and provides increased sensitivity and specificity when calling de novo variants by leveraging the parental genotype information. We then processed variant calls by using false positive filters to remove systematic artifacts. The first step uses bam-readcount (v.0.4), then identifies and flags potentially artifact variants if they fail any of the filters listed in Table S2.

Family-Relationship Testing

We evaluated familial relationships with BEAGLE's fastIBD to calculate identity by descent (IBD) between children and their parents. Variant sites within the target region in at least one family member with at least 20× coverage of the site in all individuals were used for this fastIBD calculation. If a trio failed fastIBD, defined as a parent-child pair sharing less than 40% of the target region, the trio was removed from all downstream analysis.

The SNV variant calls from the final set of 1,409 trios were combined into a multi-sample VCF file and all segregating sites were genotyped in all individuals via samtools mpileup. We then removed all sites where 50% or more individuals had a false-positive filter flag. We also removed individual variant calls with a depth (DP) less than 7 or a genotype quality (GQ) less than 20.

Table 1. Overview of Regions Sequenced and TDT Results

	Region	Candidate Gene in Region	Target Region (GRCh37)	Size (kb)	Associated SNP from GWAS	Associated SNP by Sequencing	$p_{sequence}$	Population	LD
GWAS	1p36	*PAX7*	chr1: 18,772,300–19,208,054	435.8	rs742071	rs1339062	8.88×10^{-4}	European	$r^2 = 0.726$, $D' = 0.997$
	1p22	*ARHGAP29*	chr1: 94,324,660–95,013,109	688.4	rs560426	rs560426	6.06×10^{-12}	Asian	–
	1q32	*IRF6*	chr1: 209,837,199–210,468,406	631.2	rs2013162	rs11119348	8.13×10^{-12}	Asian	$r^2 = 0.541$, $D' = 0.991$
	8q24		chr8: 129,295,896–130,354,946	1059.1	rs987525	rs7017665	8.70×10^{-11}	European	$r^2 = 0.847$, $D' = 0.983$
	10q25	*VAX1*	chr10: 118,421,625–119,167,424	745.8	rs7078160	rs10886036	8.08×10^{-9}	Asian	$r^2 = 0.941$, $D' = 0.999$
	17p13	*NTN1*	chr17: 8,755,114–9,266,060	510.9	rs9788972	rs9904526	3.07×10^{-9}	Asian	$r^2 = 0.670$, $D' = 0.999$
	17q22	*NOG*	chr17: 54,402,837–54,957,390	554.6	rs227731	rs227727	7.32×10^{-8}	Asian	$r^2 = 0.982$, $D' = 0.991$
	20q12	*MAFB*	chr20: 38,902,646–39,614,513	711.9	rs13041247	rs2865509	1.85×10^{-8}	Asian	$r^2 = 0.752$, $D' = 0.870$
Linkage	9q22	*FOXE1*	chr9: 100,357,692–100,876,841	519.1	–	rs10983654	1.31×10^{-5}	Asian	–
Candidate Gene	4p16	*MSX1*	chr4: 4,825,126–4,901,385	76.3	–	rs59138205	1.7×10^{-4}	Asian	–
	14q22	*BMP4*	chr14: 54,382,690–54,445,053	62.4	–	rs751399	0.01	Asian	–
	10q26	*FGFR2*	chr10: 123,096,374–123,498,771	402.4	–	rs12569773	1.10×10^{-4}	Asian	–
	9q22	*PTCH1*	chr9: 98,133,647–98,413,162	279.5	–	rs61076166	2.71×10^{-5}	European	–

Variants located within 75 bp of indels or dinucleotide polymorphisms occurring in >5% of samples were included in analyses but were flagged as potential artifacts. The full set of indel calls are in process and were not part of these analyses.

Variant Annotation

Variants were annotated with dbSNP137 and the Variant Effect Predictor (VEP).[13] The following VEP parameters were used: --condel b --polyphen b --sift b --hgnc --canonical. For each gene, the canonical VEP annotation was used whenever possible. In the event of multiple overlapping genes that yielded different annotations, the most damaging annotation was used. Variants were also annotated with the Combined Annotation Dependent Depletion (CADD) tool, which integrates multiple annotations (conservations, functional genomic data [e.g., DNase hypersensitivity], and protein scores [e.g., PolyPhen]) into a single score (C-score). We used the scaled C-scores, which range from 1 to 99. A score greater or equal to 10 are variants predicted to be the 10% most deleterious substitutions genome-wide. A total of 168,129 variants were identified, only 1,059 of which were coding (Figure S1).

Statistical Methods

Additional Quality Control

Prior to statistical analyses, we applied additional quality-control criteria beyond the steps outlined above. We excluded from analyses all SNPs with >5% missing genotypes over all trios. We then tested for deviation from Hardy-Weinberg equilibrium (HWE) in founders within each ancestry group (Asian and European).

Single-Variant Analysis

We tested variants with minor allele frequency (MAF) > 1% for association using the allelic transmission disequilibrium test (TDT). The TDT tests for Mendelian transmission of alleles from a heterozygous parent to the affected child. We implemented the test in PLINK,[14] which provides an allelic TDT assessing whether heterozygous parents transmit the target allele (arbitrarily defined as the minor allele) to the affected child at the expected 50% probability and generating an asymptotic p value for the χ^2 statistic (1 degree of freedom). To determine the significance threshold for tightly correlated SNPs, we applied the SimpleM method[15] to each ancestry group separately. The estimated number of independent SNPs averaged 5,200 SNPs between the Asian and European groups. Our Bonferroni p value threshold based on this number considered $p < 1 \times 10^{-5}$ to be significant, preserving a type 1 error rate of 5%.

Conditional Analysis

For variants with significant results from the single-variant TDT analyses, we performed a conditional analysis given the observed genotype at the SNP with the strongest signal to determine whether there was one signal at each region or multiple signals. This was done in a conditional logistic regression framework assumed by the genotypic TDT under a dominant model to maximize power. Conditioning was performed by fitting a multiple regression model, including the most significant marker, plus each additional individual marker in the region added sequentially. We used phased haplotype data from BEAGLE to properly account for linkage and haplotype structure in the region. We used visual inspection of both Manhattan plots and QQ plots to assess the presence of residual signal in this conditional analysis.

Analysis of Rare Variants

Rare variants (MAF ≤ 1% across all founders) were analyzed by several approaches. To select for coding variants most likely to be functional, we restricted our analysis to rare variants with a scaled CADD score of at least 10, corresponding to the top 10%

of deleterious variants genome-wide. We applied rare-variant burden tests (described below) by gene, including all variants in a gene meeting the above criteria, and restricting the analysis to sites annotated as "non-synonymous coding," "essential splice site," "stop-gained," or "stop-lost" by VEP. We took a similar approach for non-coding variants, applying burden tests to regions annotated as regulatory elements by ENCODE[16,17] or other similar efforts.[18–20] We then took an unbiased, window-based approach with ScanTrio to identify regions of significant non-random transmission of rare variants.

Rare Variant "Burden" Tests

We performed two burden-style tests for rare variants, at the gene level and for different classes of regulatory elements. The first (termed T1 here) is a binomial test comparing total numbers of transmitted and non-transmitted minor alleles for a gene/region, i.e., rare variants within the gene or region were "collapsed" prior to testing. In this analytical strategy, the observed total transmitted and non-transmitted counts are tested against a null hypothesis of equal probability of transmission and non-transmission, and p values are calculated based on a binomial distribution.

The second test (T2) uses exact probabilities calculated at each individual locus, which are then combined to obtain a joint statistic across a gene/region. More specifically, under the null hypothesis, the probability of observing the transmitted and non-transmitted minor allele counts, a_i and b_i, respectively, for the i^{th} variant is obtained based on a binomial distribution:

$$P_i = C(a_i + b_i, a_i)p_i^a(1-p)_i^b,$$

where C(n,m) is a binomial function. Once variant-specific P_i values have been obtained, they are then summarized over k variants into a logarithmized joint probability score:

$$L = \sum_{i=1}^{k} \log(P_i),$$

where L reflects the joint probability of observing all k pairs of transmission counts under the null hypothesis that none of these k variants is associated with the trait. When a_i, b_i, and k are not too large, the exact distribution of L can be obtained by enumerating the combinations of all possible a_i and b_i values under the null hypothesis, and calculating L for each combination with these formulas. Denoting $\widehat{\mathrm{L}} \ni (L_1, L_2, \ldots, L_j, \ldots \mathrm{L_M})$ as an exact distribution of L over M possible values, the p value for testing the significance of L becomes the probability of $L \leq L_j$, which can be exactly calculated as the proportion of elements of $\widehat{\mathrm{L}}$ that are less than or equal to L_j. In the situation where M is too large to be enumerated completely, a random sample can be drawn from $\widehat{\mathrm{L}}$ (with replacement) through simulation and used to approximate a p value. The advantage of the T2 statistic over T1 is that it allows both protective and deleterious rare alleles to exist within a gene or region.

We performed these tests for various subsets of rare variants, selecting based on MAF (no higher than 1% or no higher than 0.1%) and functional annotation (coding role, condel scores from VEP or CADD scores, and annotations for regulatory elements).

Scan-Trio

In addition to the gene/region based tests described above, we used the Scan-Trio method to search (via sliding windows) for sub-regions displaying over- or under-transmission of rare variants. The sliding windows were done in two ways: first considering either all possible windows of a given size or a number of markers, and second considering windows of a fixed physical distance or a fixed number of markers of overlap (to reduce the correlation in signal between adjacent, highly overlapping windows). In brief, this method involves calculating a likelihood ratio to test whether the transmission rate of minor versus major alleles differs inside the window in question, compared to the remainder of the region under consideration (comparing a 1-parameter likelihood for the observed data to a 2-parameter likelihood). In this way, we allow for either over- or under-transmission of the minor allele within a window, with the former indicating overall deleterious effect of minor alleles in the window under consideration.

Significance was assessed by permuting transmitted and untransmitted haplotypes in each region, as phased with BEAGLE, and then recalculating the Scan-Trio likelihood for each window on each permuted dataset. This approach both preserves the correlation structure between neighboring markers (i.e., the LD between nearby variants) because haplotypes are the units of permutation and allows for comparative assessment of neighboring windows with different MAFs, because significance is assessed for the same window across all permutations, thus fixing the MAFs for any given window. To identify windows showing more significant signal than was expected by chance alone, we employed QQ plots and looked for outliers from the expected distribution under the null hypothesis.

Sanger Sequencing of De Novo Mutations

We designed primers covering one or more de novo mutations with Primer3 for 82 of the 123 high-confidence de novo mutations identified in this study that were also absent from dbSNP137. The remaining de novo mutations were in highly repetitive sequence where unique PCR products could not be generated or for trios where additional samples were unavailable. PCR products for all members of the trio were sequenced on an ABI 3730XL (Functional Biosciences). Chromatograms were then transferred to a Unix workstation, base-called with PHRED (v.0.961028), assembled with PHRAP (v.0.960731), scanned by POLYPHRED (v.0.970312), and visualized with the CONSED program (v.4.0).

Follow-up Functional Studies

Zebrafish Husbandry

A pet store strain of zebrafish was maintained by standard methods except parental fish were housed at room temperature overnight prior to breeding.[21] Embryos were raised at 28.5°C. The ethical use of animals for research was approved by the University of Iowa Institutional Animal Care and Use Committee.

Zebrafish Enhancer Screen

Potential regulatory elements (i.e., *FGFR2* +254 kb [GRCh37, chr10: 123,099,588–123,100,426]; *NOG* +87 kb [GRCh37, chr17: 54,755,547–54,757,398]; and *NOG* +105 kb [GRCh37, chr17: 54,776,294–54,777,215]) were PCR amplified from human genomic DNA, cloned into the Gateway (Invitrogen) pENTR/D-TOPO vector, and transferred to the zebrafish enhancer detection vector (ZED).[22] In this dual reporter vector, one cassette is comprised of Gateway recombination sites, a minimal *gata2a* promoter, and the gene encoding enhanced green fluorescent protein (eGFP). The second cassette, which serves to report on the degree of mosaicism and also an internal control of transformation efficiency, is the cardiac-actin promoter upstream of *dsRed2* gene. The ZED vector also contains Tol2 recombination sites bracketing the entire construct to facilitate Tol2-mediated integration of the expression construct into the zebrafish genome, thus reducing cell mosaicism within the injected embryos. All plasmid constructs were sequenced and the sequence results were aligned onto the human genome with the UCSC tool BLAT[23] to ensure

fidelity of these steps. No other variants were detected within the plasmids. *tol2* mRNA was transcribed from the plasmid pKJ-Tol2.[24] ZED constructs (25–30 ng/μl) were injected along with the *tol2* mRNA (20–30 ng/μl) into 100–200 zebrafish embryos at the 1-cell stage. The developing embryos were screened at 24, 48, 72, and 96 hr after fertilization for eGFP expression. A consistent pattern of expression in a minimum of 10% of injected fish was the criterion for tissue-specific enhancer activity, which is generally sufficient to predict the expression pattern present in F1, non-mosaic transgenic lines.[25] The use of recombinant DNA was approved by University of Iowa Institutional Biosafety Committee.

Photography

Zebrafish embryos were photographed in bright field, epi-fluorescent illumination, or differential interference contrast imaging on a Leica DMRA2 compound microscope with a color 12 bit "QIClick" camera (Qimaging).

Electrophoretic Mobility Shift Assays

A full-length human *PAX7* cDNA was acquired from ATCC. The p.Ala259Val substitution was introduced into it by PCR-mediated mutagenesis.[26] cDNA encoding wild-type and p.Ala259Val PAX7 were shuttled into the CS2+ vector and the corresponding proteins were generated in vitro with a TNT kit (Promega). The products of the protein synthesis reactions were separated by gel electrophoresed NuPAGE 4%–12% Bis-Tris Gel (Life Technologies) and stained with coomassie blue (Bio-Rad). EMSA were carried out according to methods of Carey et al.[27] In brief, PAX7 proteins were incubated with 4 pmol double stranded oligonucleotide, containing the Pax7 binding site present in the *id3* promoter,[28] and end labeled with infrared tag (sequence of probe: 5′-GCTTCACCG CAATTAATGTGCATAGAGTGTGGTCACAAGATAATTCCTGA-3′). Protein and probe were incubated for 20 min at room temperature in LI-COR binding buffer, 25 μM DTT/2.5% Tween20, poly(dI-dC), sheared salmon-sperm DNA, and 50% glycerol (LI-COR). In competition experiments, an unlabeled version of the same oligonucleotide, at 10- or 100-fold excess, was added to the protein 20 min prior to addition of the labeled probe. Reaction products were electrophoresed on 4.5% poly acrylamide gel and imaged on the Odyssey Infrared Imaging System (LI-COR).

Luciferase Reporter Constructs, Transfections, and Luciferase Assays

For tests of the *PAX7* de novo mutation, a synthetic PAX7-sensitive enhancer was generated by synthesizing an oligonucleotide containing four replicates of the PAX7 binding site found in the *id3* promoter (sequence) and engineering it into pTol2-cFos-FLuc. The same *FGFR2* +254 construct used in the zebrafish studies (GRCh37, chr10: 123,099,588–123,100,426) was modified by PCR-mediated mutagenesis to test the de novo mutation at GRCh37, chr10: 123,099,960. Both variants of this element were engineered into pTol2-cFos-FLuc. This plasmid, which we generated from one described previously, contains Tol2 recombination sites, a Gateway cloning site, the c-*fos* minimal promoter, and the gene encoding firefly luciferase. Similarly constructs of the *NOG* +105 kb element (GRCh37, chr17: 54,776,294–54,777,215), containing either the major allele of rs227727 (i.e., A) or the minor allele of rs227727 (i.e., T), was engineered into pTol2-cFos-FLuc. For tests of the *PAX7* de novo mutation, HeLa cells were used, and CS2+ plasmids encoding either wild-type or the p.Ala259Val substitution were co-transfected with PAX7-sensitive reporter described above. For tests of the *FGFR2* and *NOG* non-coding elements, transient transfections were performed with Lipofectamine 3000 (Roche) into GMSM-K (human embryonic oral epithelial cells) or MC3T3-E1 (murine osteoblastic cells).

For each construct, three independent transfections were performed with Renilla luciferase (pTol2-cFos-RLuc) co-transfection, as a control for transfection efficiency. The Dual-Luciferase Reporter Assay System (Promega) and a luminometer were used to measure luciferase activity in cell lysates. All quantified results are presented as mean ± SEM. Three luciferase measurements were made on each of three independent biological replicates. A two-tailed unpaired Student's t test was used to determine statistical significance.

Cell Culture

GMSM-K human embryonic oral epithelial cell line (a kind gift from Dr. Daniel Grenier)[29] was maintained in keratinocyte serum-free medium (Life Technologies) supplemented with EGF 1-53 and bovine pituitary extract (Life Technologies). MC3T3-E1 (ATCC) murine osteoblastic cells were maintained in MEM-alpha (Life Technologies) supplemented with 10% fetal bovine serum (Life Technologies). HeLa cells were maintained in Dulbecco's Modified Eagle's Medium (Life Technologies) supplemented with 10% fetal bovine serum. All cells were incubated at 37°C and 5% CO_2.

Murine Crosses

We crossed mature C57BL/6J mice. Presence of a copulation plug on the following morning was designated as E0.5. Animal use protocols were approved by the Institutional Animal Care and Use Committees at Michigan State University.

Immunostaining

Pregnant dams were sacrificed at E13.5. Harvested embryos were fixed, embedded in paraffin, sectioned, and immunostained as described previously.[30] Primary antibodies were incubated overnight at 4°C and included NOG (Ab16054, Abcam), NTN1 (PC364, Oncogene), KRT6 (Covance, PRB-169P), p63 (Santa Cruz, 4A4, SC-8431), and IRF6 (kindly provided by Dr. Akira Kinoshita, University of Nagasaki). Secondary antibodies were incubated at room temperature for 1.5 hr and included a goat anti-rabbit (Molecular Probes, A21429) and a goat anti-mouse (Molecular Probes, A11029) antibody. Nuclei were counter-stained with DAPI (Invitrogen, D3571). We mounted all slides in ProLong Gold Antifade Reagent (Invitrogen, P36930). Imaging was performed as described previously.[30]

Mapping of Putative Transcription Factor Binding Sites

TRANSFAC 7.0, Patch 1.0, and JASPAR software were used to predict putative transcription factor binding sites within *NOG* +105 kb and *FGFR2* +254 kb elements. We used databases of mammalian and vertebrate transcription factor binding motifs for prediction of putative binding sites as previously described.[31] The binding sites were filtered based on the number of nucleotides, conservation of each nucleotide based on the consensus motif, and relevance of transcription factor.

Results

Our sequencing and analytical pipelines, depicted in Figure 1, identified potential functional variants that were de novo, common (minor allele frequency [MAF] > 1%), and rare (MAF ≤ 1%) variants in the 1,409 trios passing quality-control procedures (Table S1).

De Novo Mutations

Although de novo mutations cannot contribute directly to GWAS signals, they might help pinpoint the location of

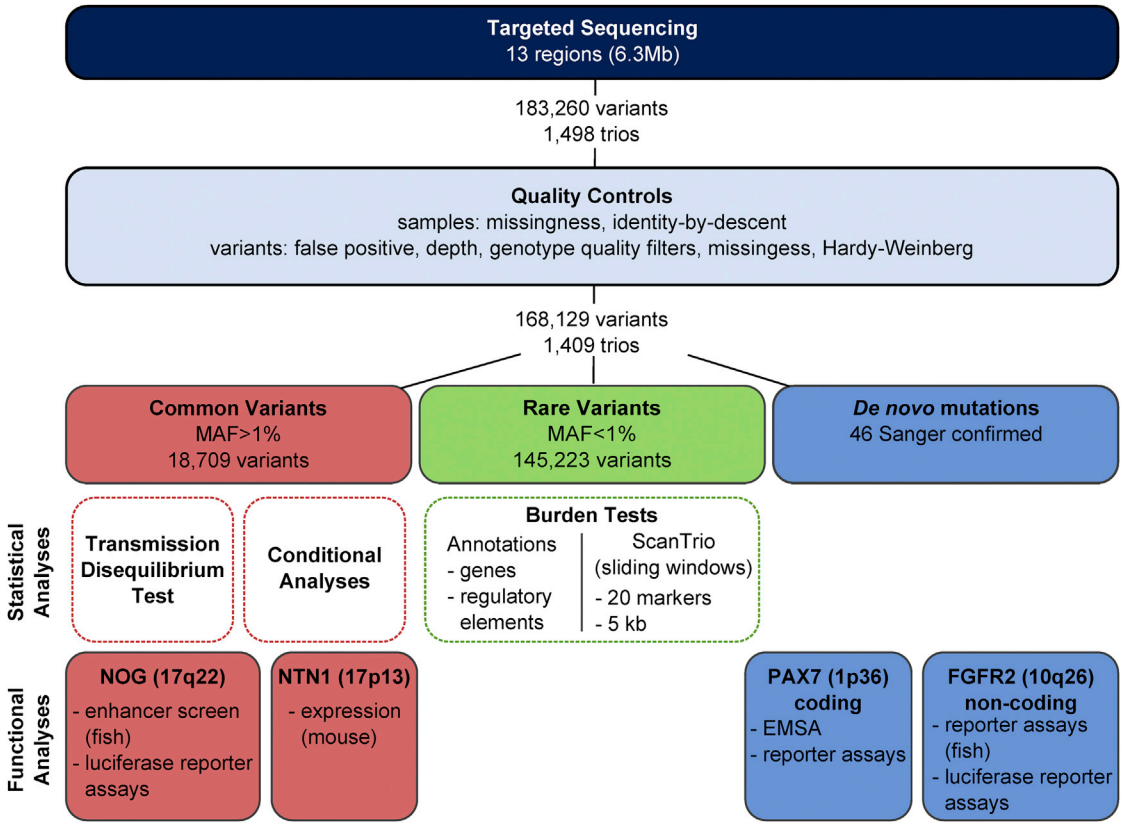

Figure 1. Diagram of Analyses Performed
Variants were categorized by allele frequency (common variants had MAF > 1%; rare variants had MAF ≤ 1%). Statistical analyses were performed and selected regions/variants were prioritized for functional analyses in vivo or in vitro.

functional common variants. We identified 123 unreported de novo mutations considered as such because they were not seen in parents and were absent from dbSNP137. One of these 123 mutations occurred in a splice-site of *IRF6* (RefSeq accession number NM_006147.3; c.1060+1G>T [p.?]). *IRF6* mutations are known to cause Van der Woude syndrome (MIM 119300), which can appear as a phenocopy of NSCL/P in ~15% of cases. Because we could not reexamine the affected child for lip pits (a defining feature of Van der Woude syndrome), this trio was excluded from further analysis in the current study. We attempted validation by Sanger sequencing for 82 de novo mutations where primers could be uniquely designed and where sufficient DNA was available; of these, 66 (80%) were confirmed (Table S3). Among these, three (3.6%) were missense mutations occurring in protein-coding genes (*ABCA4* [MIM 601691], *PIK3R5* [MIM 611397], *PAX7* [MIM 167410]). The *ABCA4* mutation was later identified in the NHLBI Exome Sequencing Project cohort and deposited in dbSNP138 (rs369860406). The *PIK3R5* mutation (RefSeq NM_014308.3; c.1085C>T [p.Ala362Val]) appeared to be mosaic because the alternate allele was found in only 36% of reads (Figure S2). We discuss the *PAX7* mutation in detail below. The majority (63, 95%) fell in non-coding intergenic or intronic DNA. We annotated these non-coding mutations for potential functional consequence using active chromatin signatures from the ENCODE project as well as recently published catalogs of craniofacial and neural crest enhancer candidates.[18,19] Eleven (16.6%) of the mutations occurred within regulatory elements annotated by Ensembl's Variant Effect Predictor (VEP) and are therefore candidates to be functional. Of these non-coding mutations, one at the *FGFR2* (MIM 176943) locus was chosen for functional analysis described below because it resides within a cranial neural crest enhancer candidate. In summary, direct sequencing of GWAS intervals led to identification of several de novo mutations within functionally annotated coding and non-coding regions.

Common Variants

To identify functional common variants, we used an allelic transmission disequilibrium test, performed conditional analysis to identify regions with multiple independent signals, annotated variants with multiple metrics, and performed follow-up functional studies. Almost all of the GWAS regions except *PAX7* showed evidence of association with p values $< 10^{-5}$ (Table 1; Tables S4 and S5; Figures S3–S10). In general, our results replicated the associations from GWASs with the variant yielding the lowest p value being either the same SNP identified by GWAS or in high LD with the reported GWAS SNP (Tables 1, S4, and S5).

Stratified analysis of the European and Asian populations showed much greater significance in the Asian trios. Only the 8q24 region showed significant associations in the European trios (rs7017665, $p = 8.7 \times 10^{-11}$). These associations were largely unchanged when the two populations were combined (Table S6). In contrast to the GWAS regions, we did not observe any significant associations near *MSX1* (MIM 142983), *FGFR2*, *BMP4* (MIM 112262), *PTCH1* (MIM 601309), or *FOXE1* (MIM 602617), all of which were selected from candidate gene studies and have not shown significance in GWASs (Figures S11–S15). Based on the conditional analysis performed, only *ARHGAP29* (MIM 610496) showed evidence for multiple independent signals, suggesting more than one common, functional variant at this locus (Figure S16). In summary, dense sequencing of GWAS regions identified additional common variants associated with NSCL/P beyond those implicated by previous GWASs.

Rare Variants

The common variants associated with NSCL/P account for only a fraction of disease heritability. The genes within GWAS regions are logical candidates to harbor rare variants, which are independent of the GWAS signal and might contribute to the risk of disease. Counts of rare coding variants per gene are listed in Table S7. To identify rare variant signals, we carried out burden tests on sets of variants based on various annotations (i.e., genes, regulatory elements; see Subjects and Methods). Neither gene nor regulatory element burden tests showed significant nonrandom transmission of rare variants after correction for multiple testing. In the ScanTrio analysis, by using an unbiased window-based approach, we experimented with different window sizes and overlaps and found signals of interest for 2 of the 13 regions (*NOG* [MIM 602991] and *NTN1* [MIM 601694], discussed in detail below).

From the results described briefly above, there were multiple regions with de novo, rare, or common variants worthy of follow-up. Comprehensive studies of these regions and variants are ongoing. Here we selected a few of the most promising regions for additional experiments by using in vitro and in vivo model systems. We selected two de novo mutations and a common variant for functional assessment. Below, we present five genes (and regions near those genes) where our statistical and/or functional analyses advanced our knowledge of the genetic etiology of CL/P.

PAX7 at 1p36

The *PAX7* region was a second-tier GWAS hit,[6] later confirmed by replication[32] and meta-analysis.[10] Although we were unable to replicate this common variant association, we identified a non-synonymous de novo mutation in *PAX7* (Figure 2A). This mutation (RefSeq NM_002584.2; c.766C>T) resulted in a substitution, p.Ala259Val, at a highly conserved residue in the DNA-binding domain (Figure 2B) and is predicted to be damaging under multiple bioinformatic algorithms (Polyphen, SIFT, CADD). Because of its location in the DNA binding domain, we hypothesized that this mutation would disrupt the ability of PAX7 to bind DNA. To test this notion we carried out electrophoretic mobility shift assays using protein synthesized in vitro and an oligonucleotide probe matching PAX7-binding regulatory sequence upstream of *ID3*.[28] In this assay, wild-type (encoding Ala259) PAX7 binds the probe more than the p.Ala259Val substitution (Figure 2C). We next carried out quantitative reporter assays in HeLa cells transfected with a luciferase reporter vector containing four copies of the PAX7 binding site. Co-transfection of a plasmid encoding wild-type PAX7 drove significantly higher expression levels relative to a plasmid encoding the p.Ala259Val substitution (Figure 2D). PAX7 is involved in neural crest induction and is expressed in cranial neural crest cells, and mice lacking *Pax7* have malformations of the nasal and maxillary structures.[33] Collectively, these results indicate that this de novo mutation disrupted PAX7 function and might contribute to CLP pathogenesis in this individual. Furthermore, they also imply that the GWAS signal in this region reflects a variant that alters the expression level or function of *PAX7* rather than another gene in the region.

ARHGAP29 at 1p22

The 1p22 locus was selected for sequencing from GWASs.[6] Our TDT results replicated the results from GWASs exactly; the peak association signal was located within an intron of *ABCA4* (rs560426, $p_{Asian} = 6.06 \times 10^{-12}$). We then extended these results by conditioning on rs560426, revealing a second signal (rs77179923, $p_{Asian} = 4.16 \times 10^{-5}$). This independent signal was located in a linkage disequilibrium block adjacent to the one containing rs560426 (Figure S16). Both peaks are located within introns of *ABCA4*, which contain regulatory elements and at least one craniofacial enhancer.[18] However, *ABCA4* is not expressed in the developing lip or palate[34] and mutations are associated with a number of ocular disorders.[35] By contrast, the neighboring gene, *ARHGAP29*, is expressed in the developing lip and palate in murine embryos.[34] Moreover, previous sequencing of *ARHGAP29* identified multiple rare variants, including a nonsense variant and a frameshift mutation, in families with NSCL/P.[34] In the present study, we identified a number of rare variants in *ARHGAP29*, including 17 previously unreported variants that in aggregate were not significantly over-transmitted to affected offspring. However, four nonsense variants (RefSeq NM_004815.3; c.976A>T [p.Lys326*]; RefSeq NM_004815.3; c.1939C>T [p.Arg647*]; RefSeq NM_004815.3; c.2367G>A [p.Trp789*]; RefSeq NM_004815.3; c.3118G>T [p.Gly1040*]) were transmitted to the affected children (we previously reported the family with the c.976A>T (p.Lys326*), variant[34]). Nonsense variants in *ARHGAP29* have never been reported by either the 1000 Genomes Project or the NHLBI Exome Sequencing Project, which cumulatively have sequenced more than

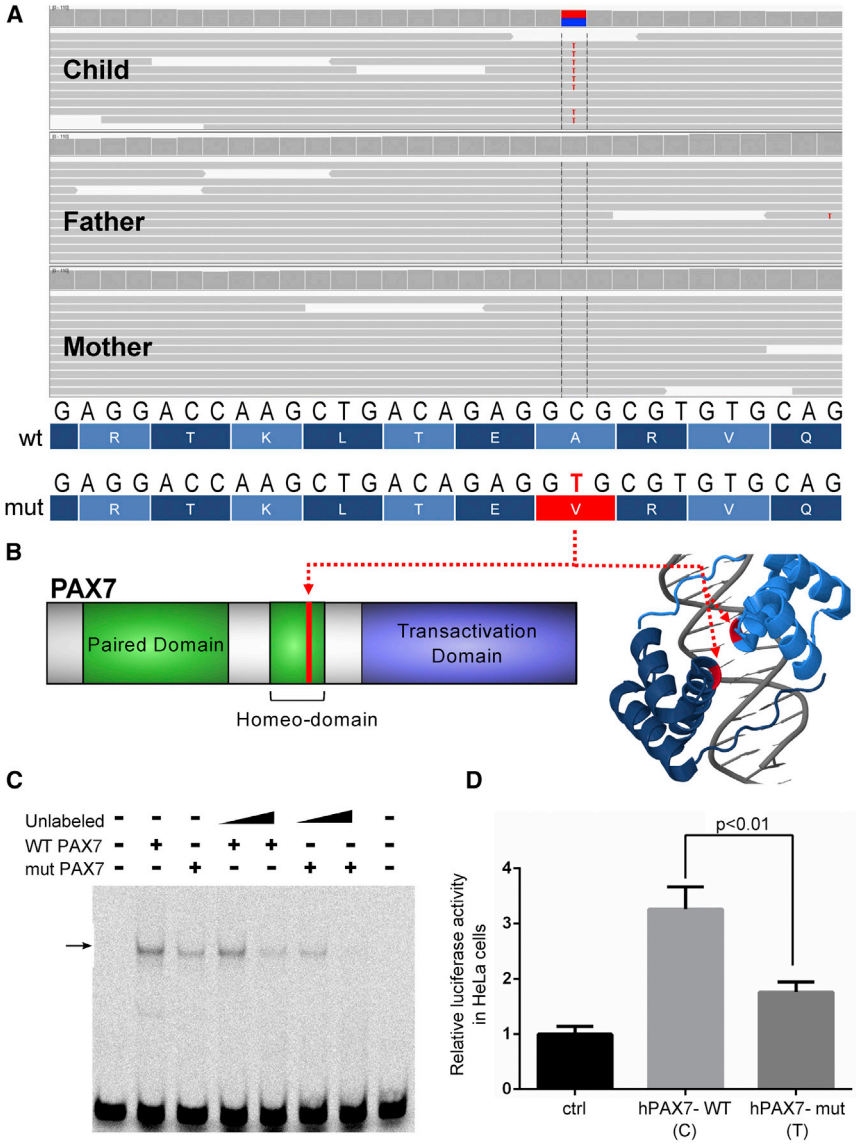

Figure 2. A De Novo Missense Mutation in *PAX7* Disrupts DNA Binding

(A) Aligned sequence reads of the affected child, father, and mother showing the de novo *PAX7* mutation, encoding p.Ala259Val, using the Integrative Genomics Viewer.

(B) Schematic of the PAX7 protein indicating the location of this mutation in the homeo-domain. Also shown is a 3D structural representation of the homeo-domain of PAX3 dimerized (dark blue and light blue ribbon structures) and in complex with DNA (gray stick representation) (Protein Data Bank ID 3CMY). PAX3 is homologous to PAX7. The position of Ala259 is marked in red and indicated by red arrows.

(C) Electrophoretic mobility shift assay using synthesized PAX7 protein. In vitro synthesized PAX7 proteins, driven by expression of the wild-type or mutant alleles as indicated, were incubated with a labeled probe containing paired and homeo-domain binding sites, based on *ID3*, or an unlabeled version of the same probes as indicated. At every dose of the competing unlabeled probe, the intensity of the band of protein-bound probe (arrow) is fainter in the reaction containing the mutant protein than in the one containing wild-type protein.

(D) Luciferase reporter assay. Cells were transfected with plasmids, encoding wild-type or mutant PAX7 as indicated, a PAX7-sensitive firefly luciferase reporter vector, and a constitutively expressed renilla luciferase reporter. Luciferase signal is the ratio of firefly and renilla luciferase measurements. Error bars: standard deviation from three replicate experiments.

7,000 individuals. Together these observations indicate that *ARHGAP29* is the gene underlying the pathogenesis of NSCL/P at this locus.

FGFR2 at 10q26

The *FGFR2* locus was selected for sequencing for several reasons: *FGFR2* plays a role in craniofacial development,[36] mutations in *FGFR2* cause two craniosynostosis syndromes that include orofacial clefting,[37,38] and rare coding variants and deletions in *FGFR2* were previously found in cases with nonsyndromic clefts.[39,40] Although neither common variants nor rare coding variants are over-transmitted to cases in our analysis, we detected a de novo mutation within non-coding DNA that possesses chromatin marks indicative of an active neural crest enhancer[17,41] (Figure 3A). This mutation (RefSeq NC_000010.10; g.123099960G>A) is located 254.6 kb downstream of the *FGFR2* transcription start site, herein referred to as the +254 kb element, and disrupts predicted transcription factor binding sites (Figure S17). In transient transgenic reporter studies in zebrafish embryos, we demonstrated that the reference allele of the human +254 kb element has enhancer activity in the neural keel (Figure 3C), brain (Figure 3D), and delaminating neural crest (Figure S18), consistent with expression of *Fgfr2* in brain and cranial neural folds in mice[42] and zebrafish. In parallel experiments, the de novo mutation revealed enhancer activity in fewer embryos (3/83) than the reference allele (41/82; $p = 1.70 \times 10^{-12}$) (Figures 3E and S18). We also tested the +254 kb element in a mesenchymal cell line in vitro and discovered that the de novo mutation had significantly lower activity than the wild-type allele (Figure 3E). These findings suggest this de novo mutation adversely affects a neural crest enhancer that, presumably, regulates *FGFR2* expression.

NTN1 at 17p13

In an earlier GWAS, the 17p13 locus was considered a second-tier hit because it did not quite reach genome-wide significance.[6] However, we found that the minor alleles

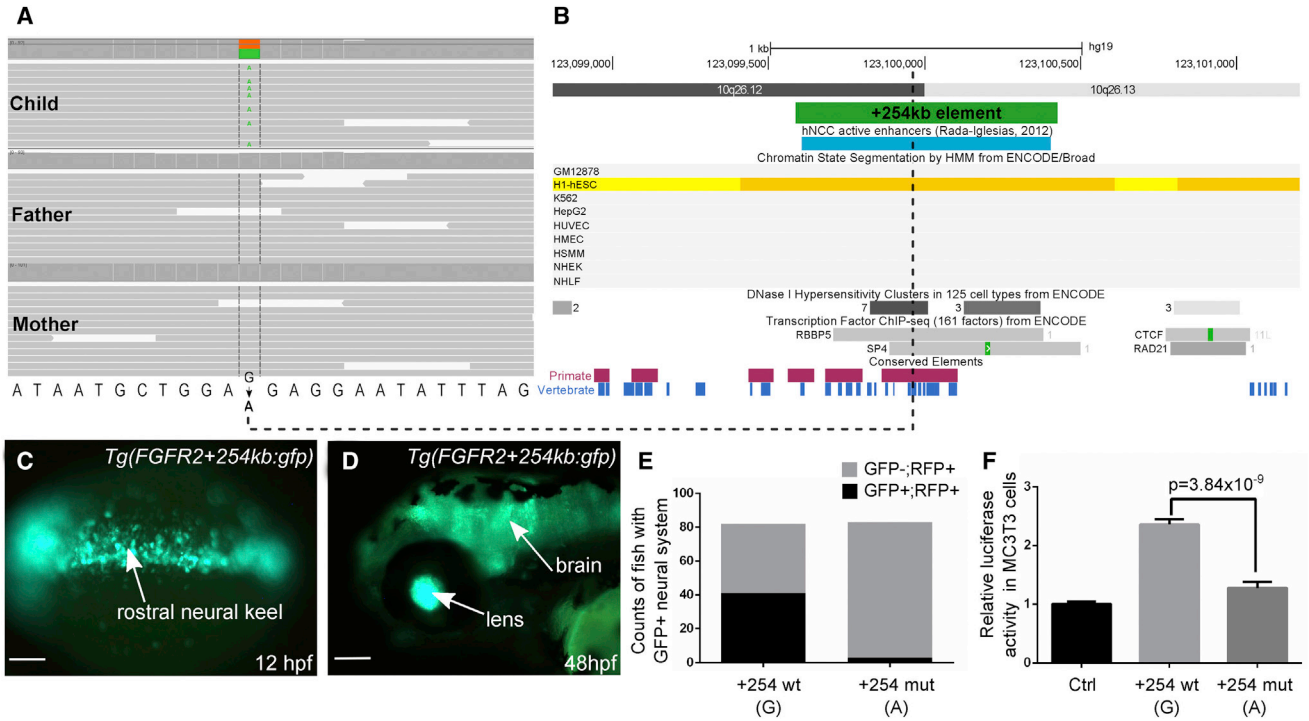

Figure 3. A De Novo Mutation in an Enhancer 254 kb Downstream of *FGFR2* Disrupts Enhancer Activity
(A) Aligned sequence reads of the affected child, father, and mother showing the de novo mutation (GRCh37: chr10: 123099960G>A).
(B) UCSC Genome Browser view of the +254 kb element. The de novo mutation falls within a human neural crest enhancer candidate defined by chromatin marks.
(C and D) Transiently transgenic embryos injected with a reporter construct containing the wild-type form of the *FGFR2* +254 kb element, a minimal promoter, and *Gfp* and imaged under epi-fluorescent illumination.
(C) Dorsal view of a 12 hr postfertilization (hpf) animal. Gfp signal is evident in the brain. Scale bars represent 20 μm.
(D) Lateral view of the head of 48 hpf animal. Gfp signal is evident in the forebrain, midrain, and hindbrain.
(E) Bar chart showing number of animals with Gfp signal in brain and RFP expression in the trunk, among animals injected with Gfp reporter vector containing wild-type and mutant forms of the *FGFR2* +254 kb element.
(F) Quantitative reporter assays in vitro comparing wild-type and de novo mutant allele of the *FGFR2* +254 kb element. MC3T3 cells were transfected with firefly luciferase vector containing wild-type or de novo mutant variants of the *FGFR2* +254 kb element, and a constitutively expressed renilla luciferase reporter. Luciferase signal is the ratio of firefly and renilla luciferase measurements. Error bars indicate standard deviation from three replicate experiments.

of multiple markers within one LD block conferred risk at genome-wide significance (e.g., rs9904526, $p_{Asian} = 3.07 \times 10^{-9}$) (Figure 4A). These markers are located between *PIK3R5* and *NTN1*. Although we cannot exclude *PIK3R5* (or other genes in this region), follow-up studies on *PIK3R5* alleles were not performed here for the following reasons. First, our ScanTrio analysis identified a protective signal in the window chr17: 8,812,274–8,830,333 (located within *PIK3R5*) containing 18 informative markers and showed a significant transmission of 6 minor alleles compared to 32 major alleles (permutation $p < 10^{-4}$) across all heterozygous rare variant genotypes in the window (Table S8). Among these variants were multiple annotations for coding variants in *PIK3R5*. Second, a homozygous missense mutation in *PIK3R5* was reported in a consanguineous family with ataxia oculomotor apraxia-3 (MIM 615217), an autosomal-recessive disorder that does not include an orofacial phenotype.[43] Finally, mice that lack *Pik3r5* are viable, whereas mice with an orofacial cleft die during the perinatal period because they cannot suckle.[44]

In contrast, *NTN1* remains a strong candidate gene. The associated common SNPs at 17p13 clustered near the transcription start site of *NTN1*. To date, no mutations in *NTN1* have been associated with any phenotype in humans. Moreover, mice that lack *Ntn1* lack the white spot of milk in the stomach and die during the perinatal period, consistent with a cleft palate phenotype.[45] To confirm that NTN1 is localized to the palate, we performed anti-NTN1 immunofluorescence on murine embryos (E13.5). Palatal shelves are composed of two epithelial layers, the periderm and the basal layer, along with underlying mesenchyme. We marked these tissue compartments by processing samples to reveal characteristic immunoreactivity: anti-KRT6 to label periderm (Figures 4B and 4C), anti-IRF6 to label both layers or oral epithelium (Figures 4D and 4E), and anti-p63 to label nuclei of the basal epithelial cells (Figures 4F and 4G). We observed high-level anti-NTN1 immunoreactivity in the mesenchyme, especially along the basement membrane of the palatal shelves (Figures 4F and 4G), and at highest levels along the presumptive medial edges and oral sides of the palatal shelves. This

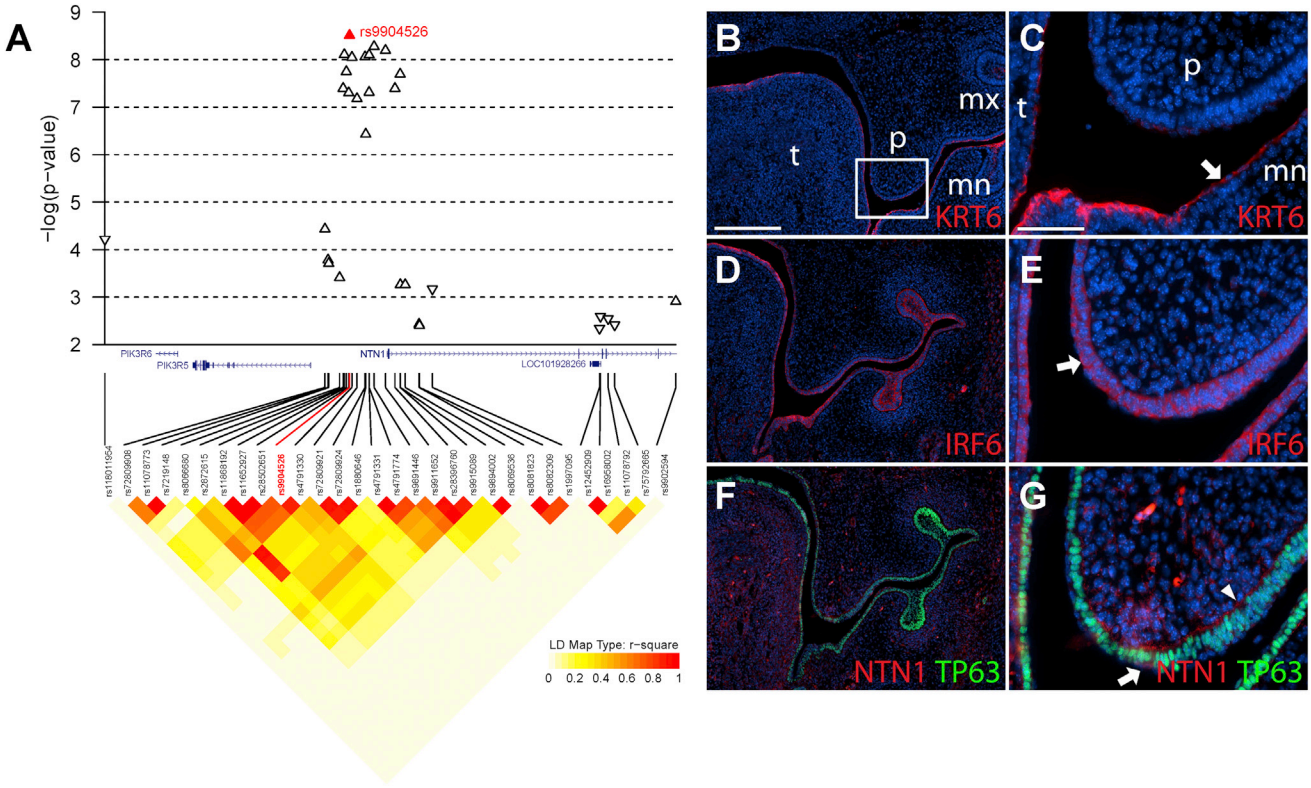

Figure 4. Association and Immunostaining of *NTN1*

(A) Regional association plot for TDT of common variants at 17p13 in 1,034 Asian trios. The SNP with the smallest p value (rs9904526) is marked by the solid red triangle. The direction of the triangles reflects the direction of association with the minor allele. Shown below the association plot are locations of genes in the region and linkage disequilibrium (measured as r^2).

(B–G) Immunostaining of coronally sectioned wild-type murine heads at E13.5. DAPI counterstains nuclei in all sections (blue). Right images (C, E, G) are magnifications or left images (B, D, F).

(B and C) KRT6 marks oral periderm (arrow) in the palate, tongue, and mandible.

(D and E) IRF6 marks the oral epithelium, including both the periderm and basal cell layer.

(F and G) NTN1 is localized to both the mesenchyme and basement membrane (arrowhead).

Scale bars represent 2 mm (B, D, F) or 50 μm (C, E, G). Labeled oral structures are the tongue (t), maxilla (mx), mandible (mn), and palatal shelves (p).

pattern is interesting because NTN1 was previously shown to regulate cell migration during embryogenesis[45–47] and in other developmental processes,[48–50] and breakdown of the basement membrane along the medial edge and invasion by mesenchymal cells is a critical step in palatal fusion.

NOG at 17q22

The 17q22 region achieved genome-wide significance in a meta-analysis of the Mangold et al.[9] and Beaty et al.[6] genome-wide scans.[10] In this region, we identified multiple SNPs reaching genome-wide significance with greatest significance detected at rs227727 ($p_{Asian} = 7.3 \times 10^{-8}$), about 105 kb downstream of the *NOG* transcriptional start site (Figure 5A). This SNP was in complete linkage disequilibrium with rs227731, the SNP with most significant association in the GWAS mentioned above. Within 1.5 kb of this common variant signal, the ScanTrio analysis of rare variants showed some evidence (permutation $p < 10^{-3}$) of combined significance of rare variants. The most significant window occurred between chr17: 54,770,168 and 54,771,787 and encompasses five informative markers, which showed overall transmission of eight minor alleles compared to no major alleles. A summary of annotation data for the variants in these regions is included in Table S8 and their location is illustrated in Figure 5B.

Prior work showed that transcripts of *Nog*, encoding a BMP antagonist, were expressed primarily in the epithelium during palatal development.[51] To further define the expression pattern, we processed mouse embryos to reveal anti-NOG immunoreactivity, again using anti-Tp63 and anti-IRF6 immunoreactivity as markers of different layers. We observed that Nog protein localized primarily to the palatal epithelium, in both basal and periderm layers, but also was detectable in mesenchyme (Figures 5C and 5D). These results are consistent with a mechanism affecting epithelial development, but because NOG is a secreted ligand, non-cell-autonomous functions are also possible. For example, overexpression of *Nog* in palatal mesenchyme caused a cell-autonomous failure of palatal shelf growth and also a non-cell-autonomous loss of epithelium.[51]

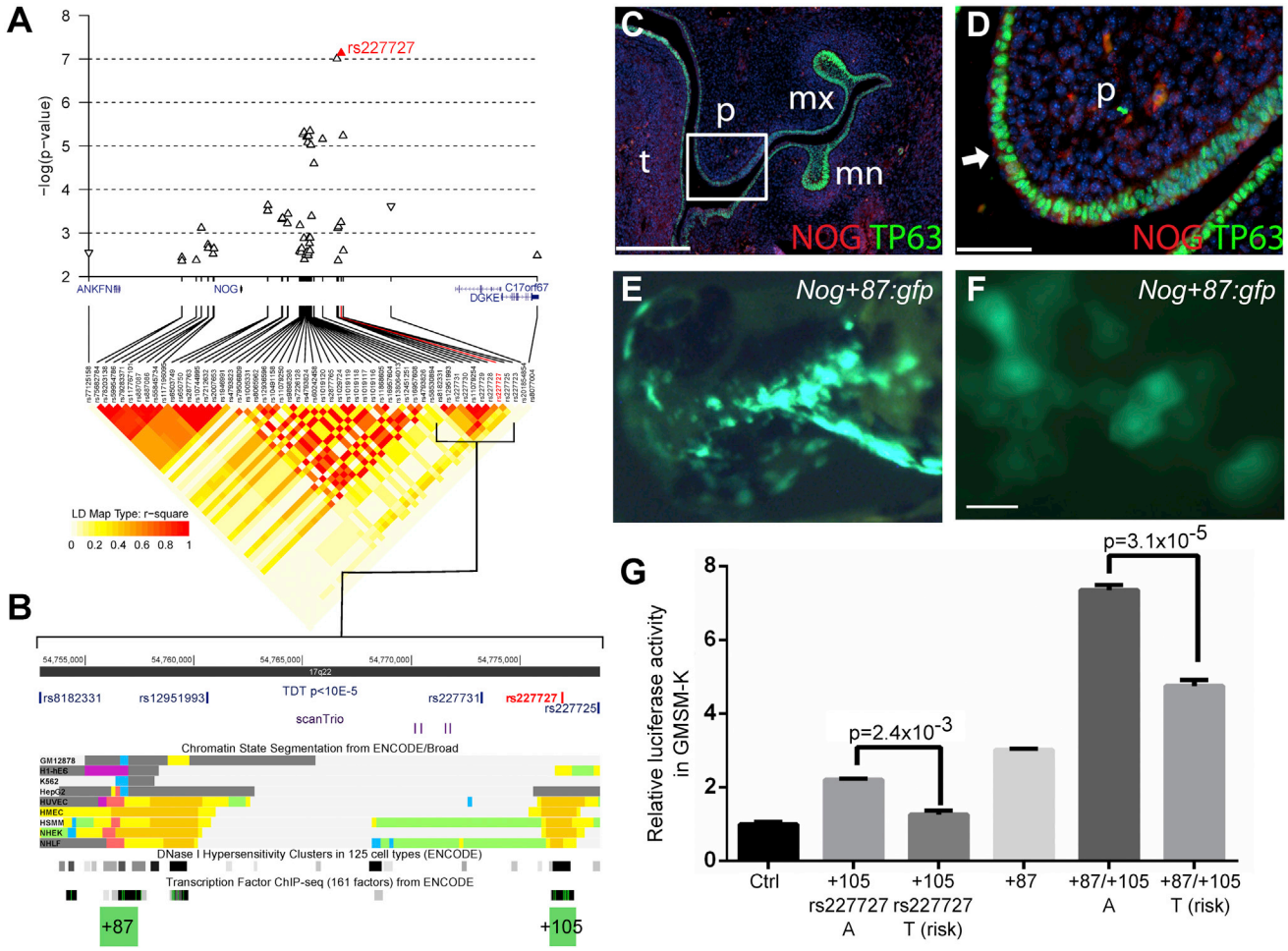

Figure 5. rs227727 Is a Functional Variant Associated with NSCL/P

(A) Regional association plot for TDT of common variants at 17q22. The SNP with the smallest p value (rs227727) is marked by the solid red triangle. The direction of the triangles reflects the direction of association with the minor allele. Shown below is the linkage disequilibrium in the region (measured as r^2).

(B) UCSC Genome Browser view of the LD block containing the most significant SNPs. In order, the tracks shown are (1) the locations of SNPs with $p < 1 \times 10^{-5}$, (2) the rare variants in the significant window from ScanTrio (in purple), (3) chromatin state segmentation tracks from ENCODE (enhancers are indicated by red, orange, and yellow bars), (4) DNase I hypersensitivity clusters from ENCODE, and (5) transcription factor ChIP-seq from ENCODE. The +87 kb and +105 kb regulatory elements are marked with green boxes.

(C and D) NOG is present in the periderm, basal cell layer, and underlying mesenchyme. Scale bars represent 2 mm (C) or 50 μm (D). Labeled oral structures are the tongue (t), maxilla (mx), mandible (mn), and palatal shelves (p).

(E and F) Transient transgenic embryos injected with a reporter construct containing the wild-type form of the *NOG* +87 kb element, a minimal promoter, and *Gfp* and imaged under epi-fluorescent illumination. GFP signal is evident in the epithelium of the 24 hpf embryos. Scale bar represents 10 μm. Surface epithelium (E) and zoomed-in view (F), showing classic hexagonal cell shape of a squamous cell epithelial cell, confirming that GFP is labeling the surface epithelium.

(G) Luciferase assay comparing the activity of *NOG* +105 kb element with wild-type or risk allele. GMSM-K cells were co-transfected firefly luciferase reporter vector with *NOG* +105 kb element with wild-type or risk allele, or tandem-engineered *NOG* +105 kb and +87 kb element, and a constitutively expressed renilla luciferase reporter. Luciferase signal is the ratio of firefly and renilla luciferase measurements. Error bars represent standard deviation from three replicate experiments.

Within the linkage disequilibrium block containing rs227727, there are two regions that are strong candidates to be active enhancers in a number of cell lines, based on the specific pattern of chromatin marks[52] (Figure 5B). Given the chromatin signature and the presence of binding sites for craniofacial transcription factors TFAP2A and c-MYC, we tested them for evidence of regulatory activity in vivo and in vitro. One element, located +87 kb from the *NOG* translation start site, is an epithelial enhancer in zebrafish (Figures 5E and 5F). The second element, at +105 kb, lacked consistent enhancer activity in transient transgenic zebrafish assays. However, this element had low-level enhancer activity in vitro in human fetal oral epithelial cells (GMSM-K) (Figure 5G) and murine osteoblastic cells (MC3T3) (Figure S19). Interestingly, the enhancer activity of the +105 kb element in both cell types was significantly lower with the NSCL/P-associated allele of rs227727 (i.e., T) than with the unassociated allele (i.e., A) (Figures 5G and S18). Because of the low-level activity of the +105 kb element in GMSM-K cells, we engineered reporter

constructs containing the +87 kb and +105 kb elements in tandem. The combined activity of these two elements is at least additive and the construct containing the rs227727 risk allele again had significantly decreased activity (Figure 5G). Of note, the disease-associated variant disrupts predicted binding sites for MEF2C and CDX2 and creates binding sites for several transcription factors, including repressors GFI1 and NKX2.5, but it is unknown which specific binding events are altered by this allele (Figure S17). In summary, we have identified a common variant, rs227727, in 100% linkage disequilibrium with the most-strongly associated SNP identified in GWAS (which now appears to be an index SNP) that alters the function of an enhancer. This is one of few examples of a common non-coding variant that is strongly associated with a structural birth defect where there is evidence that it is functional.

Discussion

To identify candidate variants that might be causal for non-syndromic cleft lip with or without cleft palate (NSCL/P), we carried out a targeted sequencing study of NSCL/P GWAS regions. After sequencing 1,409 case-parent trios from multiple Asian and European populations, we identified many candidate functional variants and carried out preliminary functional analyses on three especially strong candidates.

We identified de novo mutations in 8% of sequenced probands and further showed that two of them, one coding and one non-coding, have functional effects in our model systems. De novo mutations play a prominent role in several other complex disorders, including autism, intellectual disability, and schizophrenia.[53,54] Given the small number of coding de novo mutations identified in this study and in earlier, smaller-scale studies,[39,55] it does not appear that coding de novo mutations play a significant role in disease burden of NSCL/P. However, the contribution of de novo mutations to NSCL/P might be higher than observed here because many of the trios (40%) selected for this study came from larger multiplex pedigrees or have a reported family history of clefting; the disease in such families is unlikely to be caused by a single de novo mutation. However, even in multiplex families it is possible that de novo mutations comprise some fraction of the alleles determining the disease liability in each individual. The *PAX7* de novo mutation is only the third reported coding de novo point mutation in an individual with NSCL/P. Previous point mutations were described in *TP63*[55] and *FGF8*.[39] Of the non-coding de novo variants, we chose one located downstream of *FGFR2* as the most promising candidate and demonstrated its role in disrupting activity of a neural crest enhancer. Future investigations of the remaining de novo mutations from this study are likely to reveal additional functional effects.

About half of the heritability for NSCL/P has not been ascribed to any gene or locus, suggesting a major contribution from rare variants. Therefore it was unexpected that we observed a statistically significant over-transmission of rare variants in only 2 of 13 regions analyzed (in non-coding DNA near *NOG* and *NTN1*), notably not detecting any signal in four regions selected for sequencing based on reported contribution of rare variants (i.e., *BMP4*, *FGFR2*, *MSX1*, and *PTCH1*). Missense and nonsense mutations in *BMP4* were associated with a combination of overt cleft lip, microform clefts, and discontinuities in the superior orbicularis oris muscle.[56] However, the present study had insufficient phenotypic data to replicate this earlier result because only a small number of families in this study have undergone assessment for orbicularis oris discontinuities. Previous work on the FGF family of genes identified several interesting variants in *FGFR2* suggested to be damaging via structural protein modeling,[39] and there is evidence that *FGFR2* is associated with nonsyndromic clefting.[57] Although we did not identify a significant over-transmission of rare variants in *FGFR2*, individual variants could still prove to be functional upon further investigation. Similar conclusions pertain to *MSX1*[58] and *PTCH1*,[59] which both contain potentially damaging rare variants in our dataset, but did not show overall excess transmission of rare variants. Importantly, because functional and non-functional variants cannot be readily discerned from one another, our burden tests include both and therefore have reduced power (as seen in simulations[60]). Systematic testing of variants to identify the subset of functional variants was previously successful in other studies[61] and might be required with NSCL/P, because bioinformatics tools for predicting protein function have low accuracy. We conclude that identifying causal rare variants will require additional extensive sequencing of regions identified by GWASs or containing candidate genes in larger sample sizes and would benefit from improved algorithms for recognizing functional variants in coding and non-coding sequence.

Our TDT analyses of common variants identified strong associations in multiple regions with NSCL/P in our Asian samples, but only in a single region, 8q24, in our European trios. This was unexpected because previous candidate gene studies and GWASs identified strong associations at *IRF6*,[3,4] *FOXE1*,[11] *NOG*,[9] and *VAX1*[9] (MIM 604294) in European populations. The number of European trios sequenced might have contributed to this, because smaller numbers might create a lack of power to detect significant associations. In the present study, combining the Asian and European trios resulted in smaller p values for the associations with *NOG* and *VAX1*, indicating an additional contribution by the European trios (Table S6). However, for the *IRF6* and *FOXE1* regions, the p values were largely unchanged in the combined analysis (Tables S6). Note that in previous studies,[3,4,32,62] the significant results were driven by Northern European populations from Denmark and Norway. In contrast, our European trios are a heterogeneous group of Europeans and European Americans with self-reported white ancestry. Insufficient power

might also have contributed to the lack of association at *PAX7*, which did not reach formal genome-wide significance by GWAS until combined in a meta-analysis of the two largest studies.[10]

Nonetheless, our analyses of common variants and the surrounding regions have yielded several insights that will aid identification of pathogenic variants for NSCL/P. First, we have demonstrated how the genetic architecture of most of the sequenced GWAS regions is comparatively simple, reflected by only one common variant signal in each region. The only exception was the 1p22 region, where we identified two independent signals in the introns of *ABCA4*. Second, we identified *NTN1* as the gene likely to underlie the association at 17p13. Finally, we propose that rs227727 might itself be a functional variant at the 17q22 (*NOG*) locus and support this claim with experimental evidence. Previously, the only common variant proposed to be functional in NSCL/P was rs642961, located in an enhancer element upstream of *IRF6*.[4] In addition to rs227727, we identified numerous common variants in other regions that might demonstrate functional capabilities in further studies.

Here we demonstrated that targeted sequencing of large intervals surrounding GWAS regions is an effective approach for identifying functional rare and common variants in both coding and non-coding regions. In aggregate, our analyses highlight the important role of non-coding regulatory elements and suggest that disruption of these regions by genetic variants is a critical aspect of the pathogenesis of NSCL/P. It will be important to replicate these results in other independent populations and to sequence additional cohorts that might have unique risk alleles, for example Hispanics, Africans, and Native Americans. We conclude that sequencing of all GWAS-implicated regions in a wide range of populations, together with functional analyses, will be necessary to fully understand the role of these genes/regions in the etiology of NSCL/P.

Accession Numbers

The dbGaP accession number for the sequences reported in this paper is phs000625.v1.p1.

Supplemental Data

Supplemental Data include 19 figures and 8 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.01.004.

Acknowledgments

We are grateful to the families who participated in this study and to the staff at each recruitment site around the world, without whose efforts this study would not have been possible. We would like to acknowledge the contributions of Adela Mansilla for assistance in study design, Laura Henkle, Chika Richter, and Wendy Carricato for sample management, Nichole Nidey, Rebecca DeSensi, Margaret Cooper, and Toby McHenry for managing phenotypic data, Bhama Ramkhelawon for sharing reagents, Greg Bonde and Quynh Duong for technical assistance, the Operations Group and the Bioinformatics Group at the Genome Institute, Gabriel Sanderson for analysis pipeline support, and Holger Schwender for assistance in statistical analysis design for rare variants and for genotypic TDT. This work was supported by grants from the NIH (HG005925 [J.C.M., M.L.M.], DE008559 [J.C.M., M.L.M.], DE009886 [M.L.M.], DE016930 [M.L.M.], DE016148 [M.L.M.], DE014581 [T.H.B.], DE018993 [T.H.B.], HD073107 [R.A.C.], LM007059 [E.J.L.], GM008629 [E.J.L.], DE022696 [Y.A.K.], DE011931 [J.T.H.], HD073107 [R.A.C.]), the National Science Foundation (IOS-114722 [R.A.C.]), the Craniofacial Anomalies Research Center and the Department of Orthodontics, College of Dentistry, both at the University of Iowa (A.C.L.).

Received: November 3, 2014
Accepted: January 9, 2015
Published: February 19, 2015

Web Resources

The URLs for data presented herein are as follows:

dbGaP, http://www.ncbi.nlm.nih.gov/gap
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
ENCODE, http://genome.ucsc.edu/ENCODE/
Gene Regulation – BIOBASE Biological Databases, http://www.biobase-international.com/gene-regulation
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
JASPAR, http://jaspar.genereg.net/
OMIM, http://www.omim.org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq
ZFIN, http://zfin.org

References

1. Rahimov, F., Jugessur, A., and Murray, J.C. (2012). Genetics of nonsyndromic orofacial clefts. Cleft Palate Craniofac. J. *49*, 73–91.
2. Dixon, M.J., Marazita, M.L., Beaty, T.H., and Murray, J.C. (2011). Cleft lip and palate: understanding genetic and environmental influences. Nat. Rev. Genet. *12*, 167–178.
3. Zucchero, T.M., Cooper, M.E., Maher, B.S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., Caprau, D., Christensen, K., Suzuki, Y., Machida, J., et al. (2004). Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. N. Engl. J. Med. *351*, 769–780.
4. Rahimov, F., Marazita, M.L., Visel, A., Cooper, M.E., Hitchler, M.J., Rubini, M., Domann, F.E., Govil, M., Christensen, K., Bille, C., et al. (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat. Genet. *40*, 1341–1347.
5. Marazita, M.L., Lidral, A.C., Murray, J.C., Field, L.L., Maher, B.S., Goldstein McHenry, T., Cooper, M.E., Govil, M., Daack-Hirsch, S., Riley, B., et al. (2009). Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. Hum. Hered. *68*, 151–170.
6. Beaty, T.H., Murray, J.C., Marazita, M.L., Munger, R.G., Ruczinski, I., Hetmanski, J.B., Liang, K.Y., Wu, T., Murray, T., Fallin, M.D., et al. (2010). A genome-wide association study of cleft

lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nat. Genet. *42*, 525–529.
7. 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.
8. Grant, S.F., Wang, K., Zhang, H., Glaberson, W., Annaiah, K., Kim, C.E., Bradfield, J.P., Glessner, J.T., Thomas, K.A., Garris, M., et al. (2009). A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. J. Pediatr. *155*, 909–913.
9. 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.
10. 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.
11. Ludwig, K.U., Böhmer, A.C., Rubini, M., Mossey, P.A., Herms, S., Nowak, S., Reutter, H., Alblas, M.A., Lippke, B., Barth, S., et al. (2014). Strong association of variants around FOXE1 and orofacial clefting. J. Dent. Res. *93*, 376–381.
12. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
13. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069–2070.
14. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
15. Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., and Province, M.A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. Genet. Epidemiol. *34*, 100–105.
16. Consortium, E.P., Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., and Birney, E.; ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. *9*, e1001046.
17. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., et al. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res. *41*, D56–D63.
18. Attanasio, C., Nord, A.S., Zhu, Y., Blow, M.J., Li, Z., Liberton, D.K., Morrison, H., Plajzer-Frick, I., Holt, A., Hosseini, R., et al. (2013). Fine tuning of craniofacial morphology by distant-acting enhancers. Science *342*, 1241006.
19. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.
20. Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. Nature *461*, 199–205.
21. Westerfield, M. (1993). The Zebrafish Book (Eugene: University of Oregon Press).
22. Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., and Gómez-Skarmeta, J.L. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev. Dyn. *238*, 2409–2417.
23. Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.
24. Peng, K.-C., Pan, C.-Y., Chou, H.-N., and Chen, J.-Y. (2010). Using an improved Tol2 transposon system to produce transgenic zebrafish with epinecidin-1 which enhanced resistance to bacterial infection. Fish Shellfish Immunol. *28*, 905–917.
25. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., Urasaki, A., Kawakami, K., and McCallion, A.S. (2006). Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat. Protoc. *1*, 1297–1305.
26. Heckman, K.L., and Pease, L.R. (2007). Gene splicing and mutagenesis by PCR-driven overlap extension. Nat. Protoc. *2*, 924–932.
27. Carey, M.F., Peterson, C.L., and Smale, S.T. (2013). Electrophoretic mobility-shift assays. Cold Spring Harb Protoc *2013*, 636–639.
28. Kumar, D., Shadrach, J.L., Wagers, A.J., and Lassar, A.B. (2009). Id3 is a direct transcriptional target of Pax7 in quiescent satellite cells. Mol. Biol. Cell *20*, 3170–3177.
29. Gilchrist, E.P., Moyer, M.P., Shillitoe, E.J., Clare, N., and Murrah, V.A. (2000). Establishment of a human polyclonal oral epithelial cell line. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. *90*, 340–347.
30. Peyrard-Janvid, M., Leslie, E.J., Kousa, Y.A., Smith, T.L., Dunnwald, M., Magnusson, M., Lentz, B.A., Unneberg, P., Fransson, I., Koillinen, H.K., et al. (2014). Dominant mutations in GRHL3 cause Van der Woude Syndrome and disrupt oral periderm development. Am. J. Hum. Genet. *94*, 23–32.
31. 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.
32. Beaty, T.H., Taub, M.A., Scott, A.F., Murray, J.C., Marazita, M.L., Schwender, H., Parker, M.M., Hetmanski, J.B., Balakrishnan, P., Mansilla, M.A., et al. (2013). Confirming genes influencing risk to cleft lip with/without cleft palate in a case-parent trio study. Hum. Genet. *132*, 771–781.
33. Mansouri, A., Stoykova, A., Torres, M., and Gruss, P. (1996). Dysgenesis of cephalic neural crest derivatives in Pax7-/- mutant mice. Development *122*, 831–838.
34. Leslie, E.J., Mansilla, M.A., Biggs, L.C., Schuette, K., Bullard, S., Cooper, M., Dunnwald, M., Lidral, A.C., Marazita, M.L., Beaty, T.H., and Murray, J.C. (2012). Expression and mutation analyses implicate ARHGAP29 as the etiologic gene for the cleft lip with or without cleft palate locus identified by genome-wide association on chromosome 1p22. Birth Defects Res. A Clin. Mol. Teratol. *94*, 934–942.
35. Burke, T.R., and Tsang, S.H. (2011). Allelic and phenotypic heterogeneity in ABCA4 mutations. Ophthalmic Genet. *32*, 165–174.
36. Stanier, P., and Pauws, E. (2012). Development of the lip and palate: FGF signalling. Front Oral Biol *16*, 71–80.

37. Passos-Bueno, M.R., Serti Eacute, A.E., Jehee, F.S., Fanganiello, R., and Yeh, E. (2008). Genetics of craniosynostosis: genes, syndromes, mutations and genotype-phenotype correlations. Front Oral Biol *12*, 107–143.
38. Slaney, S.F., Oldridge, M., Hurst, J.A., Moriss-Kay, G.M., Hall, C.M., Poole, M.D., and Wilkie, A.O. (1996). Differential effects of FGFR2 mutations on syndactyly and cleft palate in Apert syndrome. Am. J. Hum. Genet. *58*, 923–932.
39. Riley, B.M., Mansilla, M.A., Ma, J., Daack-Hirsch, S., Maher, B.S., Raffensperger, L.M., Russo, E.T., Vieira, A.R., Dodé, C., Mohammadi, M., et al. (2007). Impaired FGF signaling contributes to cleft lip and palate. Proc. Natl. Acad. Sci. USA *104*, 4512–4517.
40. Osoegawa, K., Vessere, G.M., Utami, K.H., Mansilla, M.A., Johnson, M.K., Riley, B.M., L'Heureux, J., Pfundt, R., Staaf, J., van der Vliet, W.A., et al. (2008). Identification of novel candidate genes associated with cleft lip and palate using array comparative genomic hybridisation. J. Med. Genet. *45*, 81–86.
41. 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.
42. Orr-Urtreger, A., Givol, D., Yayon, A., Yarden, Y., and Lonai, P. (1991). Developmental expression of two murine fibroblast growth factor receptors, flg and bek. Development *113*, 1419–1434.
43. Al Tassan, N., Khalil, D., Shinwari, J., Al Sharif, L., Bavi, P., Abduljaleel, Z., Abu Dhaim, N., Magrashi, A., Bobis, S., Ahmed, H., et al. (2012). A missense mutation in PIK3R5 gene in a family with ataxia and oculomotor apraxia. Hum. Mutat. *33*, 351–354.
44. Suire, S., Condliffe, A.M., Ferguson, G.J., Ellson, C.D., Guillou, H., Davidson, K., Welch, H., Coadwell, J., Turner, M., Chilvers, E.R., et al. (2006). Gbetagammas and the Ras binding domain of p110gamma are both important regulators of PI(3)Kgamma signalling in neutrophils. Nat. Cell Biol. *8*, 1303–1309.
45. Serafini, T., Colamarino, S.A., Leonardo, E.D., Wang, H., Beddington, R., Skarnes, W.C., and Tessier-Lavigne, M. (1996). Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. Cell *87*, 1001–1014.
46. Salminen, M., Meyer, B.I., Bober, E., and Gruss, P. (2000). Netrin 1 is required for semicircular canal formation in the mouse inner ear. Development *127*, 13–22.
47. Srinivasan, K., Strickland, P., Valdes, A., Shin, G.C., and Hinck, L. (2003). Netrin-1/neogenin interaction stabilizes multipotent progenitor cap cells during mammary gland morphogenesis. Dev. Cell *4*, 371–382.
48. Park, K.W., Crouse, D., Lee, M., Karnik, S.K., Sorensen, L.K., Murphy, K.J., Kuo, C.J., and Li, D.Y. (2004). The axonal attractant Netrin-1 is an angiogenic factor. Proc. Natl. Acad. Sci. USA *101*, 16210–16215.
49. van Gils, J.M., Derby, M.C., Fernandes, L.R., Ramkhelawon, B., Ray, T.D., Rayner, K.J., Parathath, S., Distel, E., Feig, J.L., Alvarez-Leite, J.I., et al. (2012). The neuroimmune guidance cue netrin-1 promotes atherosclerosis by inhibiting the emigration of macrophages from plaques. Nat. Immunol. *13*, 136–143.
50. Ramkhelawon, B., Hennessy, E.J., Ménager, M., Ray, T.D., Sheedy, F.J., Hutchison, S., Wanschel, A., Oldebeken, S., Geoffrion, M., Spiro, W., et al. (2014). Netrin-1 promotes adipose tissue macrophage retention and insulin resistance in obesity. Nat. Med. *20*, 377–384.
51. He, F., Xiong, W., Wang, Y., Matsui, M., Yu, X., Chai, Y., Klingensmith, J., and Chen, Y. (2010). Modulation of BMP signaling by Noggin is required for the maintenance of palatal epithelial integrity during palatogenesis. Dev. Biol. *347*, 109–121.
52. Consortium, E.P.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
53. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., et al. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863–885.
54. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. *43*, 585–589.
55. Leoyklang, P., Siriwan, P., and Shotelersuk, V. (2006). A mutation of the p63 gene in non-syndromic cleft lip. J. Med. Genet. *43*, e28.
56. Suzuki, S., Marazita, M.L., Cooper, M.E., Miwa, N., Hing, A., Jugessur, A., Natsume, N., Shimozato, K., Ohbayashi, N., Suzuki, Y., et al. (2009). Mutations in BMP4 are associated with subepithelial, microform, and overt cleft lip. Am. J. Hum. Genet. *84*, 406–411.
57. Wang, H., Zhang, T., Wu, T., Hetmanski, J.B., Ruczinski, I., Schwender, H., Liang, K.Y., Murray, T., Fallin, M.D., Redett, R.J., et al. (2013). The FGF and FGFR gene family and risk of cleft lip with or without cleft palate. Cleft Palate Craniofac. J. *50*, 96–103.
58. Jezewski, P.A., Vieira, A.R., Nishimura, C., Ludwig, B., Johnson, M., O'Brien, S.E., Daack-Hirsch, S., Schultz, R.E., Weber, A., Nepomucena, B., et al. (2003). Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. J. Med. Genet. *40*, 399–407.
59. Mansilla, M.A., Cooper, M.E., Goldstein, T., Castilla, E.E., Lopez Camelo, J.S., Marazita, M.L., and Murray, J.C. (2006). Contributions of PTCH gene variants to isolated cleft lip and palate. Cleft Palate Craniofac. J. *43*, 21–29.
60. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. Eur. J. Hum. Genet. *21*, 1158–1162.
61. Davis, E.E., Zhang, Q., Liu, Q., Diplas, B.H., Davey, L.M., Hartley, J., Stoetzel, C., Szymanska, K., Ramaswami, G., Logan, C.V., et al.; NISC Comparative Sequencing Program (2011). TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. Nat. Genet. *43*, 189–196.
62. Moreno, L.M., Mansilla, M.A., Bullard, S.A., Cooper, M.E., Busch, T.D., Machida, J., Johnson, M.K., Brauer, D., Krahn, K., Daack-Hirsch, S., et al. (2009). FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. Hum. Mol. Genet. *18*, 4879–4896.