



Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21

Nandita Mukhopadhyay¹ · Madison Bishop² · Michael Mortillo³ · Pankaj Chopra² · Jacqueline B. Hetmanski⁴ · Margaret A. Taub⁵ · Lina M. Moreno⁶ · Luz Consuelo Valencia-Ramirez⁷ · Claudia Restrepo⁷ · George L. Wehby⁸ · Jacqueline T. Hecht⁹ · Frederic Deleyiannis¹⁰ · Azeez Butali¹¹ · Seth M. Weinberg^{1,12} · Terri H. Beaty⁴ · Jeffrey C. Murray¹³ · Elizabeth J. Leslie² · Eleanor Feingold^{1,12,14} · Mary L. Marazita^{1,12,15}

Received: 26 September 2019 / Accepted: 7 December 2019 / Published online: 17 December 2019
© The Author(s) 2019

Abstract

Orofacial clefts (OFCs) are among the most prevalent craniofacial birth defects worldwide and create a significant public health burden. The majority of OFCs are non-syndromic, and the genetic etiology of non-syndromic OFCs is only partially determined. Here, we analyze whole genome sequence (WGS) data for association with risk of OFCs in European and Colombian families selected from a multicenter family-based OFC study. This is the first large-scale WGS study of OFC in parent–offspring trios, and a part of the Gabriella Miller Kids First Pediatric Research Program created for the study of childhood cancers and structural birth defects. WGS provides deeper and more specific genetic data than using imputation on present-day single nucleotide polymorphic (SNP) marker panels. Genotypes of case–parent trios at single nucleotide variants (SNV) and short insertions and deletions (indels) spanning the entire genome were called from their sequences using human GRCh38 genome assembly, and analyzed for association using the transmission disequilibrium test. Among genome-wide significant associations, we identified a new locus on chromosome 21 in Colombian families, not previously observed in other larger OFC samples of Latin American ancestry. This locus is situated within a region known to be expressed during craniofacial development. Based on deeper investigation of this locus, we concluded that it contributed risk for OFCs exclusively in the Colombians. This study reinforces the ancestry differences seen in the genetic etiology of OFCs, and underscores the need for larger samples when studying for OFCs and other birth defects in populations with diverse ancestry.

Introduction

Orofacial clefts, primarily cleft lip (CL) and cleft palate (CP) are among the most common birth defects in all populations worldwide with differences in birth prevalence by ancestry (Dixon et al. 2011; Rahimov et al. 2012). Surgical treatment along with ongoing orthodontia, speech and other therapies is very successful in ameliorating the physical

health effects of OFC, but there is still a significant social, emotional and financial burden for individuals with OFC, their families, and society (Nidey et al. 2016; Wehby and Cassell 2010). Furthermore, there are disparities in access to such therapies for OFCs (Nidey and Wehby 2019), similar to other malformations with complex medical and surgical needs. Some studies have suggested a reduced quality of life for individuals with OFCs (Naros et al. 2018), while other studies have identified higher risk to certain types of cancers (Bille et al. 2005; Bui et al. 2018; Taioli et al. 2010). Thus, it is critical to identify etiologic factors leading to OFCs to improve diagnostics, treatments, and outcomes.

The causal genes for most syndromic forms of OFCs are now known, and listed within OMIM (<https://www.omim.org/search/advanced/geneMap>, search term = (cleft lip cleft palate syndrome) AND “omim snp”[Filter]), but the majority of OFC cases—including about 70% of CL with or without CP (CL/P) and 50% of CP alone—are considered

Elizabeth J. Leslie, Eleanor Feingold and Mary L. Marazita are co-senior authors.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00439-019-02099-1>) contains supplementary material, which is available to authorized users.

✉ Mary L. Marazita
marazita@pitt.edu

Extended author information available on the last page of the article

non-syndromic, i.e. they occur as isolated anomalies with no other apparent cognitive or structural abnormalities (Dixon et al. 2011). The causal genes for non-syndromic OFCs are still largely undiscovered. To date, there have been 52 genome-wide associations reported and replicated between non-syndromic CL/P and genetic markers (NHGRI-EBI Catalog of published genome studies) (Buniello et al. 2019), but as for most other complex human traits (Hazelett et al. 2016; Tak and Farnham 2015; Zhu et al. 2017), very few putative functional variants for non-syndromic OFCs have been identified from genome-wide association studies (GWASs) (Beaty et al. 2016). In particular, the high heritability for OFC, estimated at 90% by a twin study in a Danish sample (Grosen et al. 2011) cannot be explained by all identified common variants significantly associated with OFC, sometimes referred to as the “missing heritability” problem (Manolio et al. 2009). Additional approaches will be necessary to expand our understanding of genetic variation in non-syndromic OFCs and whole genome sequencing (WGS) holds the promise of teasing out the so-called missing heritability from GWASs of OFC and other complex traits (Wainschein et al. 2019).

An important new approach has been implemented by the Gabriella Miller Kids First Pediatric Research Consortium (<https://commonfund.nih.gov/kidsfirst/overview>). Kids First was established in 2015 to address gaps in our understanding of the genetic etiologies of structural birth defects and pediatric cancers by providing WGS of case–parent trios with these major pediatric conditions. Addressing both of these areas (structural birth defects and pediatric cancers) in Kids First was partially motivated by the observation that children with birth defects such as OFCs are at a higher risk of also developing some cancers, and their family members also have elevated risk (Bille et al. 2005; Bui et al. 2018), suggesting there may be shared genetic pathways underlying cancer and birth defects. The KidsFirst study consists of 952 case–parent trios (i.e. affected probands and their parents) from multiple OFC studies, of which, 415 are of European descent, 275 Latino, 125 Asian and 137 African. The current study summarizes initial findings on common variants, i.e. single nucleotide polymorphic (SNP) markers and small insertions/deletions from WGS of a sample of 315 trios European descent, as well as a sample of 265 trios of Latin American ancestry from Colombia, all with offspring affected with CL/P.

Methods

Study design

Two samples of case–parent trios were analyzed for the current family-based association study, one of European

descent recruited from sites around the United States, Argentina, Turkey, Hungary and Madrid, and a second of trios from Medellin, Colombia. The two samples are referred to as **European** and **Colombian**, respectively, in this study. Recruitment of participants and phenotypic assessments were done at regional treatment centers for orofacial clefts after review and approval by the site-specific IRBs (see Ethics Statement at the end of “Methods”).

This study included case–parent trios consisting of affected offspring and their parents (Table 1). Most of the **European** parents and all **Colombian** parents are unaffected for CL/P (see breakdown of trios in Table 1). All trios had offspring with a cleft lip or a cleft lip plus cleft palate, and had not been diagnosed with any recognized genetic syndrome. The affection status was defined as cleft lip with or without cleft palate (CL/P) for all analyses here because the Colombian sample did not have the breakdown between cleft lip alone (CL) versus cleft lip with cleft palate (CLP). Table 1 shows the counts of GMKF trios sequenced for the present study, by their country of origin.

Genetic data

Whole genome sequencing of the European sample was carried out at the McDonnell Genome Institute (MGI), Washington University School of Medicine in St. Louis, while sequencing of the Colombian sample was conducted at the Broad Institute, both with an average of 30× coverage. Variant calling on the European trios was performed using pipelines at MGI, and aligned to the GRCh37/hg19 genome assembly. The European sample’s genotypes were realigned and recalled by the GMKF’s Data Resource Center at Children’s Hospital of Philadelphia to match the Colombian sample, which was aligned to hg38 and called using GATK pipelines (DePristo et al. 2011; McKenna et al. 2010; Van der Auwera et al. 2013) at the Broad Institute (<https://software.broadinstitute.org/gatk/best-practices/workflow>). The alignment and joint genotyping workflow used to

Table 1 Counts of CL/P trios by recruitment site and cleft type

Sample	Total Trios	Trios with no affected parents	Trios with 1 affected parent	Trios with 2 affected parents
European	315	280	32	3
Site: USA	209	185	21	3
Hungary	56	51	5	
Madrid	30	26	4	
Argentina	1	1		
Turkey	19	17	2	
Colombian	265	265		
Total	580	545	32	3

harmonize these two samples of case–parent trios was developed using GATK Best Practice recommendations, with the goal of being functionally equivalent with other current large genomic research efforts. Briefly, the harmonization pipeline first converted the mapped alignments within each sample to unmapped alignments, then re-ran the GATK genotyping workflow, namely base quality score recalibration (BQSR), simultaneous calling of SNPs and indels using single-sample variant calling (HaplotypeCaller), multiple-sample joint variant calling, and finally refinement and filtering of called variants. Data processing and storage of harmonized results were done on the Cavatica platform within an Amazon Web Services (AWS) environment. The GMKF Data Resource Center (DRC) was responsible for tracking, final checking, and release of the variant calls via its portal. The released variant data contained genotypes called at 35,600,754 single nucleotide variants (SNVs) and 4,320,146 indels mapped to the hg38 reference sequence. Details of the harmonization process are provided “Kids First DRC Genomics Harmonization Pipeline Description” in the supplement.

Assessment of sample data quality and data cleaning

Each sample of trios (**European, Colombian**) was separately analyzed for genotyping inconsistencies, at an individual level, as well as on a trio basis. *Genotype quality* Genotypes with either unacceptable read depth (minimum depth 10 reads for autosomes; minimum 5 reads for X chromosomes in males), or genotyping quality (minimum GQ 20; minimum GQ = 10 for X chromosome variants in males) were first set to unknown. *Sample quality* Each individual’s set of variant calls was checked for excess heterozygosity (> 3 standard deviations from mean heterozygote/homozygote ratio), deviant transition to transversion ratios ($T_s/T_v > 3$ standard deviations from mean T_s/T_v across samples), low genotyping rates (below 90%), and for inconsistency between the average homozygosity on the X-chromosome and the individual’s reported sex. Each trio was assessed for Mendelian error rates and deviation from the expected degree of relatedness between each set of parents and offspring. Genomes flagged for sex or relationship issues were compared with SNP array genotypes from the POFC Multiethnic study (Leslie et al. 2016) to resolve sample swaps or misclassification of sex, where possible (some trios from our study were not part of POFC Multiethnic study). A trio was excluded if it failed more than one of these data quality tests, and if recovery was not possible after comparison with the SNP array genotype data.

After QC procedures, the final dataset consisted of 315 complete European trios and 265 complete Colombian trios. Biallelic variants including SNPs and short indels (indels range between 1 and 10,000 BP in length) with a genotyping

rate of at least 90% were included in our analyses. A total of 5,374,579 variants were analyzed in the **European** trios, and 4,905,638 in **Colombian** trios. Of these, 4,220,712 variants were analyzed for the **Combined** trios.

Genome-wide association testing of SNPs and indels

Genome-wide association was conducted using two versions (allelic and genotypic) of the transmission disequilibrium test (TDT), for each polymorphic variant. The PLINK software (Chang et al. 2015; Purcell and Chang 2019) was used to run the standard genome-wide allelic TDT (aTDT), which does not consider the parents’ cleft status. We also ran genotypic TDT (gTDT) (Schaid 1996) on the trios, and compared the association of p values with those from the aTDT. Effect sizes are not directly comparable between the two methods. The aTDT compares the transmission of a target allele to the affected child from heterozygous parents (Spielman et al. 1994), and is based on McNemar’s Chi-squared statistic. Because only heterozygous parents can contribute to this statistic, statistical power is greatly influenced by minor allele frequency (MAF) and one population may have considerably more or less power at any given SNP when MAF varies across populations. The gTDT compares the observed genotype in the child to “pseudo-controls” representing other genotypes possible from the parental mating type. Schwender et al. (2012, 2014) demonstrated an efficient method for computing this gTDT statistic. Because either TDT represents a test of strict Mendelian inheritance of the marker (despite sampling case–parent trios through the affected proband), this test is robust to spurious associations arising from population stratification and can provide greater power for rare phenotypes (Laird and Lange 2008). The null hypothesis of either TDT is the complete absence of either linkage between the marker and an unobserved causal gene or linkage disequilibrium (LD) between the marker and an unobserved causal gene. Rejection of this composite null hypothesis implies the presence of both linkage and LD. The TDT is most appropriate for our study, given our participants originate from diverse populations, and the Colombians in particular are known to reflect varying degrees of admixture of African, Hispanic, Native American and European genes.

Three genome-wide TDT analyses were run: separately in European and Colombian trios and then in all trios combined. Significant p values for the allelic TDT statistic were calculated using the exact binomial distribution. Although the TDT statistic is robust to population substructure, an overall TDT analysis can mask subgroup specific results, thus principal component analysis (PCA) was run on the parents separately for each sample (**European, Colombian**) and the normalized eigenvalues were examined for evidence of sub-groups within each sample. For PCs producing eigenvalues exceeding ± 5 , we conducted genetic association

assuming an additive model using the eigenvalues of each individual as quantitative traits. The PCA was conducted using the KING program (Manichaikul et al. 2010). PLINK (Chang et al. 2015; Purcell and Chang 2019) was used to run the quantitative association.

Identification of significant associations

Due to our limited sample sizes, only SNPs with a minor allele frequency of at least 10% within each sample of trios were considered in these TDT analyses. The allelic TDT test relies on asymptotics, and can give inflated associations for lower MAF SNPs at this sample size when applied genome wide. We subsequently examined lower MAF SNPs in specific regions for fine-mapping purposes (see below). The genome-wide threshold for significant association was set at $5.0e-08$, and the critical value for suggestive association was set at $1.0e-05$.

Fine-mapping and rare-variant association in 21q region

A subset of the genome-wide significant associations (i.e. those not overlapping with previously reported OFC genes/regions) was selected for more in-depth investigation. All biallelic variants with a genotyping rate of 90% or greater, regardless of MAF, were investigated within each region of interest (defined as 1 Mb centered on each lead variant). Each interval was annotated for possible roles in craniofacial development by literature searches of all genes contained within that interval, functional annotation of variants using multiple tools including Bystro (Kotlar et al. 2018), Variant Effect Predictor (McLaren et al. 2016), and HaploReg (Ward and Kellis 2016). We also queried the UCSC genome browser's gene-by-gene interaction track for known OFC genes/regions. This track identifies genes reported in protein interaction databases and recognized biological pathways (Poon et al. 2014).

Rare variant (RV) association using the TDT framework was run only for regions containing SNPs showing significant evidence of linkage and association in the aTDT. For each association peak, we identified all genes located within 500 KB of the lead SNP, and selected non-synonymous RVs within the exons of these genes. Burden and collapsing methods were used, as our dataset is composed solely of case–parent trios, and these tests were applied to each gene separately, after phasing the observed genotype data of common SNPs. Beagle was used to calculate haplotypes (Browning and Browning 2007) using all variants within a selected region. The RV-TDT software (He et al. 2014) was then run on phased haplotypes for exonic, non-synonymous SNVs in genes with a minimum of four variant sites. RV-TDT reports burden and combined multivariate and collapsing

(CMC) types of test statistics, as well as a weighted sum statistic. The observed MAFs within European and Colombian parents were used to calculate weights for each RV, where SNVs with smaller MAFs receive higher weights. Some of the RV-TDT statistics use phased haplotypes to calculate empirical p values by permuting the haplotypes of each parent. In addition to the exonic rare variants, we also selected intronic and intergenic variants and analyzed these using RV-TDT. Intronic and intergenic variants were divided into subsets based on gene locations in this region, and analyzed using a procedure similar to the exonic, non-synonymous SNPs.

Ethics Statement

University of Pittsburgh IRB served as the coordinating center for the entire study with approval under IRB#91-06-10-03-2. Written consent was obtained from subjects through site-specific IRBs at the University of Pittsburgh, University of Iowa and Johns Hopkins University (HRPO#03-0871, IRB#HSC-MS-03-090, IRB#970405, IRB#200109094, IRB#200109094).

Results

Genome-wide association of SNPs and indels

Genome-wide associations using allelic and genotypic transmission disequilibrium test (TDT) were run separately in 315 **European** and 265 **Colombian** trios and then in the **Combined** set of all 580 trios on bi-allelic single nucleotide polymorphic (SNP) markers and indels with minor allele frequency (MAF) greater than 10% (see “**Methods**” for discussion of MAF cutoff). A comparison of the p values between allelic TDT (aTDT) and genotypic TDT (gTDT) showed high concordance (see “**Comparison between aTDT and gTDT**” in Supplementary and Supplementary Figure S1, Fig. 1); therefore, only the aTDT results are discussed in the following sections. p values calculated using the exact binomial distribution from McNemar's test are reported for the aTDT.

Tables 2 and 3 show the most significant results in the **European** (Table 2) and **Colombian** trios (Table 3). Several SNPs gave genome-wide significantly associated p values in the stratified aTDT analysis of **European** (Table 2 and Fig. 1 top panel) and **Colombian** trios (Table 3 and Fig. 1 middle panel), and a single SNP achieved genome-wide significance in the **Combined** sample (Fig. 1 bottom panel). In the **European** sample, 17 significant associations are observed across multiple chromosomes (Table 2). In the **Colombian** sample, four significant associations are observed for markers on chromosomes 6, 8, 19 and 21. After close examination of the

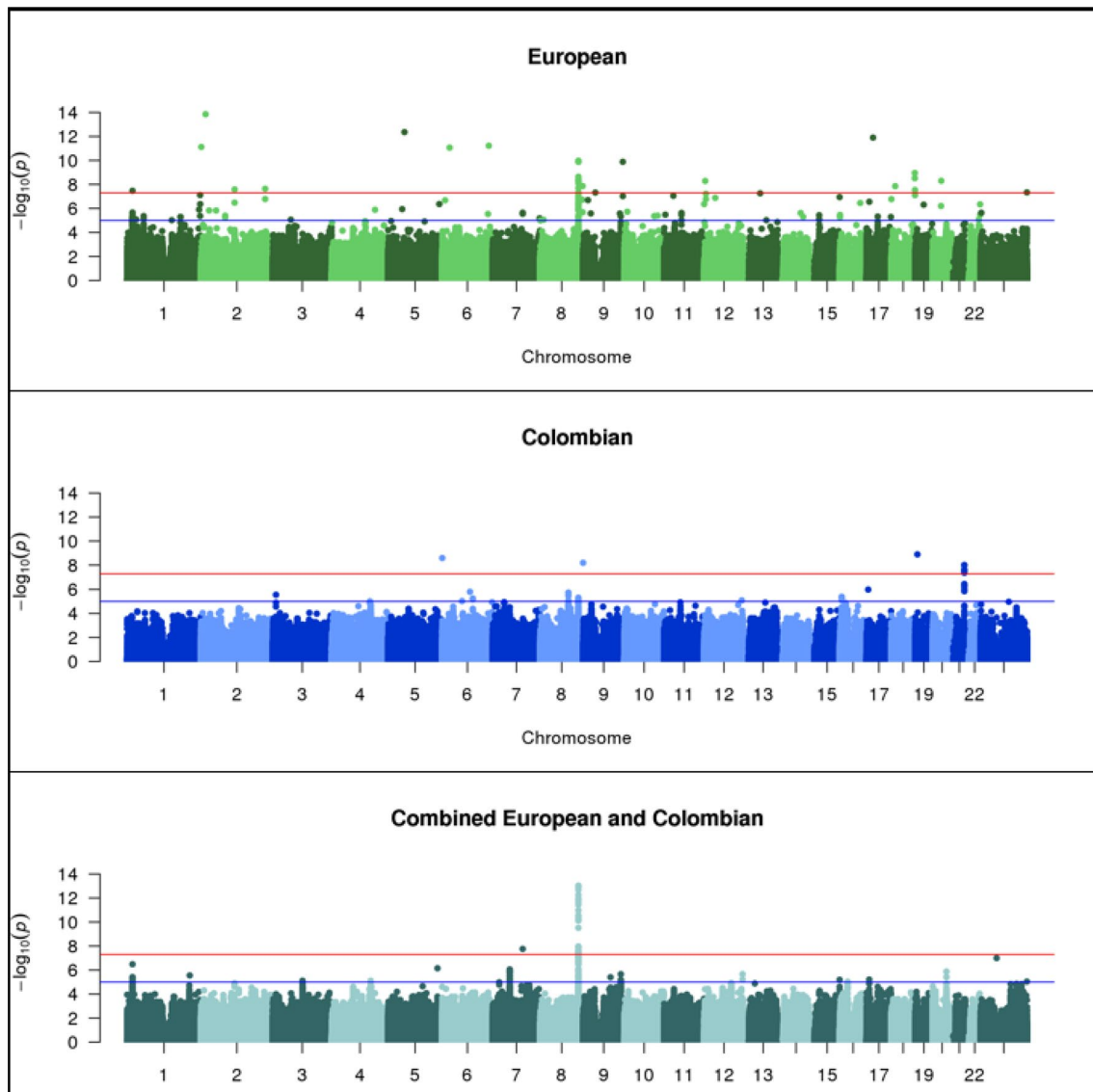


Fig. 1 Manhattan plots of European (315 trios), Colombian (265 trios) and Combined (580 trios) allelic TDTs

genome-wide significant associations in the **European** and **Colombian** trios, the one strongly supported new result was a region on chromosome 21q22.3, discussed below. In the **Combined** aTDT, a single genome-wide significant association ($p=9.35E-14$, $OR=2.13$, $95\% CI=[1.74-2.62]$, SNP rs72728755) was observed in the 8q24.21 chromosomal region. Many of the other associations showed properties that reduced our confidence in their reliability, which included (1) no additional variants yielding either significant or suggestive p values close to the lead SNP, (2) the lead SNP was located in a highly repetitive region, or (3) the lead SNPs showed substantial differences in MAF across European or Latino samples in gnomAD (Karczewski et al. 2019). Therefore, we concluded that these might not be reliable signals. Note that the first criterion alone was not

sufficient to make us deem a result unreliable, as the 10% MAF cutoff may have been responsible for single-SNP association peaks.

Comparison between allelic TDTs of European and Colombian trios

A qualitative comparison of the **European** and **Colombian** aTDT results showed few commonalities between the two analyses of common SNPs. Except for the peaks at the 8q24.3 chromosomal region, all other genome-wide significant regions in the **European** trios were neither significant nor suggestive in the **Colombian** trios, and vice versa. The lack of new signals from the **Combined** trios supports this observation. For the purposes of comparison, Table 2 lists

Table 2 Significant associations in **European** (315 trios) compared with **Colombian** (265 trios) and **Combined** (580 trios)

Significantly associated locus in European aTDT	RS number (bp position) of lead variant in European aTDT	<i>p</i> value (effect size) of lead variant in European aTDT	Strongest association seen near European lead variant in Colombian aTDT		Strongest association seen near European lead variant in Combined aTDT	
			<i>p</i> value (OR)	RS number (bp position)	<i>p</i> value (OR)	RS number (bp position)
1p36.13	rs78998514 (18,608,118)	3.4E–08 (2.05)	2.2E–04 (1.83)	rs753305 (18,143,515)	9.2E–06 (1.55)	rs78998514 (18,608,118)
2p25.3	rs1362227148 (1,361,834)	7.6E–12 (0.32)	5.0E–04 (0.51)	rs13429476 (968,756)	7.1E–04 (0.67)	rs72762992 (907,551)
2p24.3	rs36094286 (15,787,755)	1.4E–14 (0.13)	2.7E–04 (1.71)	rs7569215 (16,017,189)	1.2E–03 (1.42)	rs340727 (16,207,847)
2q14.1	chr2:113,497,779	2.6E–08 (0.31)	6.3E–03 (0.65)	– (113,537,068)	7.1E–05 (1.81)	rs112243068 (113,381,134)
2q35	rs1164161401 (216,293,984)	2.3E–08 (0.22)	4.2E–05 (1.52)	rs3770473 (216,634,116)	8.9E–05 (1.74)	rs2712179 (216,768,013)
5q11.2	rs1290483247 (54,785,929)	4.4E–13 (0.13)	3.4E–04 (0.54)	rs113820400 (54,451,286)	9.9E–04 (0.65)	rs113820400 (54,451,286)
6p22.2	rs1747567 (25,529,642)	8.6E–12 (0.22)	1.8E–02 (1.64)	rs9366622 (25,414,309)	4.7E–04 (1.49)	rs34164888 (25,521,693)
6q25.3	chr6:157,311,140	5.98E–12 (0.33)	3.83E–03 (0.53)	rs9505843 (157,522,349)	8.6E–04 (1.61)	rs34164888 (157,582,486)
8q24.21	rs72728755 (128,978,136)	1.29E–10 (2.39)	4.92E–06 (2.37)	rs79382561 (128,819,668)	1.4E–14 (2.13)	rs72728755 (128,978,136)
8q24.3	rs1429661747 (143,179,754)	1.4E–08 (0.31)	2.7E–03 (1.89)	rs57681929 (143,410,437)	3.6E–03 (0.71)	rs7463227 (143,187,836)
9p11.2	rs1471353675 (40,816,247)	4.8E–08 (0.37)	2.7E–03 (1.65)	– (41,288,651)	1.2E–01 (0.79)	– (41,155,200)
9q34.2	rs879409092 (133,278,859)	1.3E–10 (0.08)	2.5E–03 (1.89)	rs2073921 (133,162,643)	5.6E–04 (0.66)	rs62576050 (133,525,936)
12p13.32	rs1293776695 (3,555,780)	5.2E–09 (0.24)	1.4E–04 (0.57)	rs727864 (3,307,233)	1.3E–04 (1.48)	rs588106 (3,122,022)
12p13.31	rs1463969293 (5,928,511)	6.0E–08 (0.20)	9.6E–04 (1.58)	rs61917137 (6,260,869)	3.3E–03 (1.35)	rs216852 (5,975,025)
17p11.2	rs1446333119 (21,895,128)	1.3E–12 (0.11)	6.4E–03 (0.70)	rs8080056 (21,545,419)	9.9E–04 (0.73)	rs8080056 (21,545,419)
18p11.21	rs576835177 (13,288,784)	1.4E–08 (0.21)	3.4E–04 (0.56)	rs12957953 (13,180,059)	1.5E–03 (1.36)	rs11080665 (13,643,180)
18q23	rs1381043271 (79,225,853)	1.1E–09 (0.25)	2.7E–03 (0.64)	rs11876371 (79,778,636)	8.5E–04 (0.69)	rs11876371 (79,778,636)
20q11.1	rs1321001584 (29,360,893)	5.0E–09 (0.25)	8.8E–04 (0.46)	28937230 (28,937,230)	2.9E–02 (0.79)	– (29,801,022)
Xq28	rs306890 (155,757,485)	4.6E–08 (2.09)	3.2E–03 (1.49)	rs145079381 (155,955,827)	2.0E–04 (1.62)	rs150716120 (155,757,485)

p values reported for **Colombian** and **Combined** trios are located within 500 MB of the lead SNP in the **European** trios

all **European** peaks and contains the least associated *p* values with their corresponding estimated odds ratios (OR) observed in the **Colombian** and **Combined** aTDTs within 500 KB on either side of each **European** peak SNP (Table 2 columns 4–7). Since allele frequencies for specific SNPs may differ between the two samples, this provides a region-level view of replication across the samples. Similarly, Table 3 lists the **Colombian** peaks, along with the minimally associated *p* values and corresponding odds ratios observed in the **European** and **Combined** aTDTs within 500 KB on

either side of each **Colombian** peak. As seen in Tables 2 and 3, **European** and **Colombian** trios differ considerably with respect to the genomic regions that show significant association with CL/P.

Previously reported OFC risk loci

Two of the genome-wide significant associations observed in this study, 1p36.13 and 8q24.21, have been previously reported as associated with risk to OFCs by our group and

Table 3 Significant associations in **Colombian** (265 trios) compared with **European** (315 trios) and **Combined** (580 trios)

Significantly associated locus in Colombian aTDT	RS number and bp position of lead variant in Colombian aTDT	<i>p</i> value (effect size) of lead variant in Colombian aTDT	Strongest association seen near Colombian lead variant in European aTDT		Strongest association seen near Colombian lead variant in Combined aTDT	
			<i>p</i> value (OR)	RS number (bp position)	<i>p</i> value (OR)	RS number (bp position)
6p25.3	rs376150594 (677,242)	2.6E−09 (0.27)	4.2E−04 (0.51)	rs62389424 (422,631)	2.4E−04 (0.60)	rs59342393 (900,025)
8q24.3	rs879371667 (144,767,652)	6.4E−09 (0.28)	1.5E−02 (0.73)	rs2979293 (144,965,922)	2.0E−02 (1.33)	rs2730064 (144,743,132)
19p13.2	rs113870866 (7,692,010)	1.3E−09 (0.11)	5.2E−05 (1.69)	rs74176226 (7,296,552)	1.3E−03 (0.67)	– (7,794,108)
21q22.3	rs2839575 (42,706,006)	9.8E−09 (2.48)	1.8E−04 (0.45)	– (42,629,765)	1.2E−05 (1.62)	rs2839575 (42,706,006)

p values reported for **European** and **Combined** trios are located within 500 MB of the lead SNP in the **Colombian** trios

others (Beaty et al. 2010; Birnbaum et al. 2009; Ludwig et al. 2012). The 1p36.13 peak is located 23 kb upstream of the transcription start site of the *PAX7* gene. These associations were significant only in our **European** trios, consistent with previous studies suggesting a stronger association in participants of European ancestry compared to other racial/ethnic groups (Leslie et al. 2015).

The 8q24.21 region has been consistently implicated in nearly all previous OFC studies especially among samples of European ancestry. The lead SNP among **Europeans** (rs55658222) is in strong linkage disequilibrium (LD) with another SNP rs987525 in the HapMap European sample. The rs987525 SNP was found to be the lead SNP in this region in several previous GWASs and also showed modest evidence of association and linkage in the **Colombian** trios (*p* value 8.609e−06, odds ratio = 1.984, CI = [1.46–2.69]). In the **European** trios, a suggestive association was observed for an indel located at 9,295,770 bp on chromosome 17, approximately 52 kb centromeric to the *NTN1* gene (*p* = 2.77e−07, odds ratio = 0.29, CI = [0.18–0.48]). No other previously reported OFC variant reached even a suggestive level of significance (suggestive threshold *p* < 1.0e−05) in our WGS study, which is not unexpected given the smaller sample size of this WGS study compared to published GWASs. Supplementary Table S2 shows the most significant aTDT *p* values within 500 KB of all previously reported OFC risk variants.

Chromosome 21q22.3 association in the Colombian trios

We observed genome-wide significant associations in the **Colombian** trios within a 30 kb interval on chromosome 21q22.3 (Fig. 2, top panel). In this sample, the common variants had relatively large estimated odds ratios ranging from 2.33 to 2.48, i.e. approximately twofold increases in the transmission of the risk alleles from parents to the proband

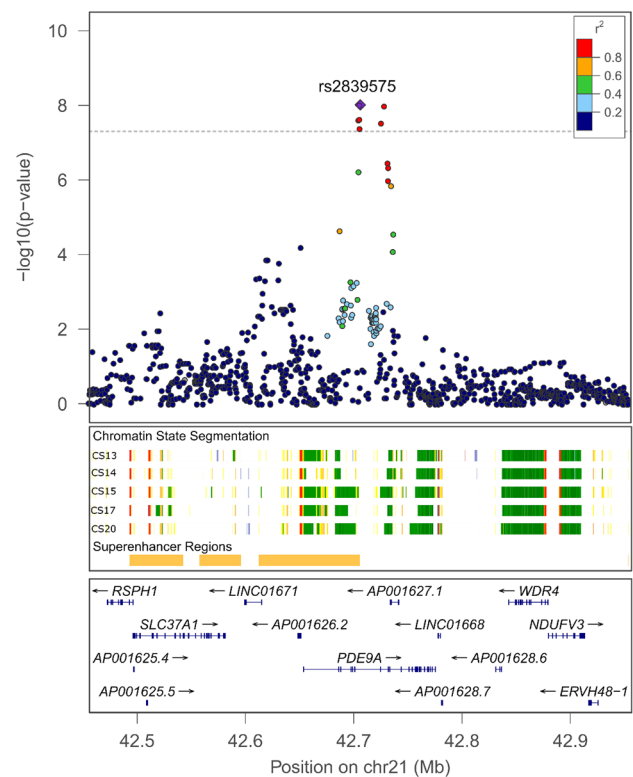


Fig. 2 Regional plot of $-\log_{10}(p)$ for SNPs and indels in chromosome 21q22.3 peak region

offspring. The smallest *p* value was observed at rs2839575 (*p* = 9.75e−09, odds ratio = 2.48, 95% CI = [1.81–3.45]).

GWAS of a Latino sample from a previous study, the POFC Multiethnic study, reported suggestive association at this genomic region [see Fig. 1 in Leslie et al. (2016)]. That Latino sample included diverse Hispanic groups from the US, Guatemala, Argentina, and Colombia, and all of the current WGS *Colombia* trios. However, the POFC

Multiethnic study also had 129 additional Colombian trios. In that study, the GWASs of Asian and European samples did not show association in this region, nor did the combined GWAS of all the POFC Multiethnic study samples. The fact that the current WGS case–parent trio study yielded a genome-wide significant association with a smaller sample size suggests this association might be unique to **Colombians**. We explored the validity and implications of this observation through a number of analyses, as described below.

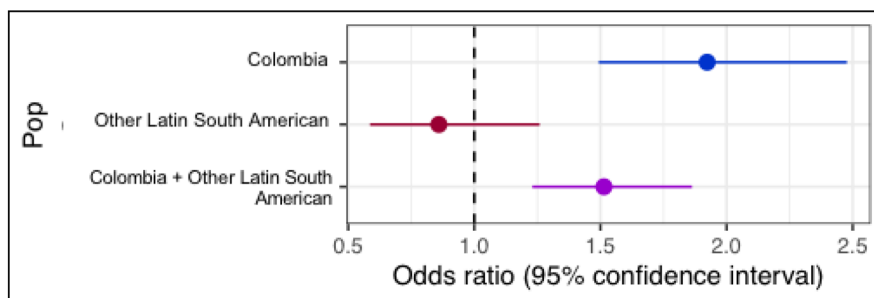
We first examined the aTDT p values for our **Colombian** WGS trios using their SNP array data from the POFC Multiethnic study. The p values in this region were nearly identical to those observed in our WGS association, confirming the association we observed here was not an artifact of sequencing.

We next investigated whether population substructure within the Colombian parents could have caused the observed association in the WGS data by examining the ancestry principal components (PCs) as well as results of quantitative association between PCA eigenvalues to variants within the peak region (see “Methods” for details). PCA showed no evidence of population substructure (Supplementary Figure S1, Fig. 2a), and no association was observed between the eigenvalues and variants in the chromosome 21q22.3 region (Supplementary Figure S1, Fig. 2b). A

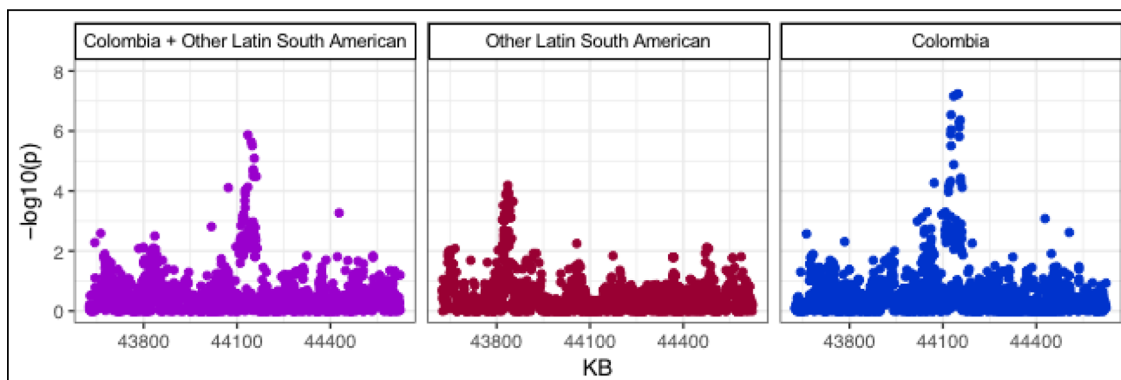
positive association between eigenvalues and variants would have indicated that the observed association with CL/P is in reality due to population substructure; therefore, this association did not appear to be an artifact of population admixture.

We verified that this region does not show evidence of association in other Latin Americans, by reanalyzing imputed genotype data of independent Latino trios from the previously published POFC Multiethnic GWAS study (Leslie et al. 2016). The aTDT p value and corresponding odds ratio at rs2839575 observed in the Colombian subjects were considerably different from those in the Latino sample, and the non-Colombian Latino trios showed no significant association at rs2839575 (Fig. 3a, forest plot). Moreover, the combined set of non-Colombian Latinos resulted in much weaker associations across a 1 MB region flanking SNP rs2839575 as well as for this SNP itself. The odds ratios at the rs2839575 variant showed an opposite (although non-significant) effect in the non-Colombian Hispanics as compared to Colombians (Fig. 3b, regional p value plot and Supplementary Table S3). We concluded from the stratified aTDT results that this SNP influences OFC risk only in Colombians.

We, therefore, investigated the possibility of ancestry differences between our Colombian sample and the other Latino populations. Ancestry principal components



(A) Odds ratio at rs2839575 in Colombia and other Latin South American samples



(B) TDT p values in a 1 MB region surrounding rs2839575, in Colombia and other Latin South American samples

Fig. 3 Estimated odds ratios (with 95% CI) and $-\log_{10}(p)$ values from the aTDT in Colombia and other Latino samples

calculated from the POFC Multiethnic SNP genotype data (unrelated individuals only) showed Colombians to be ancestrally diverse from the other Latino populations (Supplementary Figure S1, Fig. 3).

Given that the 21q22.3 association is observed only in the Colombian sample and that the ancestry of Colombians is different from the other Latin American samples, we checked whether the absence of an association signal in the other Latin American samples merely reflects differences in MAF rather than differences in true effects of risk alleles. That is, it is possible that a causal variant exists in all populations but has a considerably higher frequency (or is in LD with a variant of higher frequency) in the Colombians. Given the population history of Colombians, causal OFC variants may have arisen from one particular ancestral group, and such variants may be more frequent (and therefore more informative) among Colombians. The origin of African ancestry of Colombians is different from that of the other Latino populations (Gouveia et al. 2019). We, therefore, looked at the frequencies of the Colombian risk alleles across different populations. For this analysis, we again turned to genotyped and imputed SNP genotypes from the POFC Multiethnic study. The MAFs of the 30 most significantly associated SNPs within the 21q22.3 peak region in Colombian trios were compared to 15 populations defined by country of recruitment from the POFC Multiethnic study. None of these 30 SNPs had higher MAF among Colombians compared to other Latino populations (Supplementary Figure S1, Fig. 4). Moreover, the 15 most significant SNPs in this peak region had higher allele frequencies in all other population groups (European, African, and Asian) compared to Colombians or other Latinos. Thus, there was no conclusive evidence that population-specific variants contributed to the association signal seen in this study. However, several of these variants had estimated odds ratios between 1.1 and 1.5 in Asian, Europeans, or Africans, suggesting these variants in this region may also increase risk for OFCs in other populations, but at a reduced level.

Finally, we tested for effects of rare variants within the **Colombian** trios using burden and collapsing tests because we observed a number of low-frequency and rare variants with large odds ratios in this region (see “[Methods](#)” for rare variant testing procedure). Common variants with the strongest associations were all intronic variants within the *PDE9A* gene; however, all had moderate odds ratios around 2.0. In this region, there were 37 SNPs with minor allele frequencies near or below 1% in the **Colombian** trios and estimated OR > 5 (Supplementary Table S4), including mainly intronic and a few intergenic SNVs (28 intronic, 8 intergenic). The exception was one non-synonymous SNV, rs138007679 in the *RSPHI* gene (aTDT odds ratio 8, 95% CI = [1.001–63.96]), which produces an amino acid change (A > C, leucine to tryptophan according to ClinVar). Alone,

this variant does not clearly implicate *RSPHI* over other genes in the region, so we performed a rare variant TDT on all non-synonymous variants within the 13 genes falling in this region. None of the individual genes achieved the nominal significance (Supplementary Table S5), so this result remained inconclusive. We also carried out rare variant TDTs of intronic and intergenic variants with similar results, finding only nominally significant associations attributable to intergenic, low-frequency variants (MAFs ranging between 0.5 and 1%).

In the absence of any clearly pathogenic variant or gene based on combined effects of rare variants, we examined regulatory elements and protein–protein interaction pathways in this region with respect to craniofacial development. All associated variants below a suggestive level of significance ($p < 1.0e-05$) were located within the *PDE9A* gene, which does not have any known role in controlling risk to OFCs. However, the *PDE9A* gene overlaps a super-enhancer region for craniofacial development identified from histone profiling in early human craniofacial development (Wilderman et al. 2018). Multiple genes in the region, including *PDE9A*, appear to be actively transcribed during human craniofacial development (Fig. 2). Another gene of interest is *UBASH3A*, located ~220 kb centromeric to this peak signal. The UBASH3A protein was previously shown to physically associate with *SPRY2* via a yeast two-hybrid assay (33). *SPRY2* has been reported by GWASs of OFC and shown required for palatogenesis in mice (Welsh et al. 2007); whether UBASH3A is also expressed in craniofacial structures has not yet been determined.

Discussion

This study is the first large-scale WGS study of OFCs, one of the most common birth defects worldwide, using a case–parent trio design. We conducted association analyses of common variants from WGS in two samples of case–parent trios, one of European ancestry and the other of Latin American ancestry from Colombia. We replicated two known OFC loci and identified a promising new region on chromosome 21 in the Colombian sample. A combined association analysis of these two samples together clearly shows that OFC risk loci differ by ethnicity. The 8q24 locus has been repeatedly shown to be associated with risk of OFCs in both case–control and case–parent trio samples from a range of ethnicities such as Europeans and Latin Americans, with some evidence from Asians (Murray et al. 2012). Here, we found slight differences in the larger 8q24 region between Europeans and Latin Americans but there appears to be a shared risk locus at 8q24.21, consistent with Colombians having a strong influence from European ancestry. *IRF6*, a gene that has been linked to OFCs in samples of Asian and Latino

ancestry, was not detected in our Colombian trios, possibly due to the small sample size.

We observed evidence of linkage and association with a previously unreported region on chromosome 21 spanning the *PDE9A* gene only in the Colombian sample. We verified that this locus is unique to Colombians, by running separate aTDTs in Colombian and non-Colombian Latino trios using imputed genotype data from the previous POFC Multiethnic GWAS study (Leslie et al. 2016). We examined whether the apparent risk alleles have ancestral origins from non-Latino populations and noted that the estimated effect sizes were slightly elevated in Asian, European and African populations although never achieving genome-wide significance. However, larger or more phenotypically specific samples may be necessary to find conclusive statistical evidence. The significantly associated common variants in the chromosome 21q22.3 peak were mostly intronic or intergenic, with no obvious biological function. There were a number of rare variants with large aTDT odds ratios, including a non-synonymous SNP within the *RSPH1* gene; however, TDT of rare coding non-synonymous variants did not provide conclusive statistical evidence of association between genes in this region and CL/P. Although none of the genes in this region are known to contribute to the development of OFCs, they appear to be actively transcribed during human craniofacial development and should be examined further in follow-up studies.

Our study is limited by the actual sample sizes of the two sets of trios. The effective sample sizes for testing association using the TDT may be reduced further as not all trios are informative at any given variant. This may have impacted our ability to replicate associations from prior CL/P GWASs. Genetic heterogeneity would also reduce power to detect association. The analysis of ancestry PCAs indicated the existence of genetic heterogeneity in both samples. While the TDT test ensures the validity of our association peaks in the presence of genetic heterogeneity, the power to detect association is adversely affected. For example, subjects from Spain and Turkey are distinct from the rest of the **European** cohort, and quantitative association of PC eigenvalues of **European** trio-parents shows significant association in the vicinity of the 8q24 locus. In conclusion, a few previously reported CL/P loci were replicated by our study, and a new plausible CL/P locus was observed. Larger, more homogeneous samples would be needed to verify the other significant associations seen in our study and detect causal rare variants.

Acknowledgements These studies are part of the Gabriella Miller Kids First Pediatric Research Program consortium (Kids First), supported by the Common Fund of the Office of the Director of the National Institutes of Health (www.commonfund.nih.gov/KidsFirst). Washington University's McDonnell Genome Institute was awarded an administrative supplement (3U54HG003079-12S1) to sequence structural birth defect samples including the European descent

Orofacial Cleft samples for the current study funded through Kids First (X01-HL132363, MLM, EF). Further, the Broad Institute Sequencing Center was awarded a grant (U24-HD090743) to sequence structural birth defect cohort samples including the Latin American Orofacial Cleft family samples for the current study funded through Kids First (X01-HL136465, MLM, EF). The sequencing centers plus the Kids First Data Resource Center (kidsfirstdrc.org, supported by the NIH Common Fund through U2CHL138346) provided quality control analyses in support of this project. The data analyzed and reported in this manuscript were accessed from dbGaP [www.ncbi.nlm.nih.gov/gap; **European trios:** dbGaP accession number phs001168.v2.p2; **Colombian trios:** dbGaP accession number phs001420.v1.p1] and from the Kids First Data Resource Center (kidsfirstdrc.org). Additional grants supported the assembling of the cohorts, collection of the phenotypic data and samples, and data analysis [R01-DE016148 (MLM, SMW), R03-DE026469 (MLM, EF), R01-DE012472 (MLM), U01-DD000295 (GLW), R01-DE014581 (THB), R01-DE011931 (JTH), R21-DE016930 (MLM), and R01-DE014667 (Andrew Lidral and LMM), and R00-DE025060 (EJL)]. Many thanks to the participating families and study teams worldwide without whom these studies would not have been possible. Additional thanks to Andrew Lidral, Mauricio Arcos-Burgos, and Andrew Czeizel.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Beaty TH 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. <https://doi.org/10.1038/ng.580>
- Beaty TH, Marazita ML, Leslie EJ (2016) Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Research* 5:2800. <https://doi.org/10.12688/f1000research.9503.1>
- Bille C, Winther JF, Bautz A, Murray JC, Olsen J, Christensen K (2005) Cancer risk in persons with oral cleft—a population-based study of 8093 cases. *Am J Epidemiol* 161:1047–1055. <https://doi.org/10.1093/aje/kwi132>
- Birnbaum S et al (2009) IRF6 gene variants in Central European patients with non-syndromic cleft lip with or without cleft palate. *Eur J Oral Sci* 117:766–769. <https://doi.org/10.1111/j.1600-0722.2009.00680.x>
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association


- studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084–1097. <https://doi.org/10.1086/521987>
- Bui AH, Ayub A, Ahmed MK, Taioli E, Taub PJ (2018) Association between cleft lip and/or cleft palate and family history of cancer: a case-control study. *Ann Plast Surg* 80:S178–s181. <https://doi.org/10.1097/sap.0000000000001331>
- Buniello A et al (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47:D1005–d1012. <https://doi.org/10.1093/nar/gky1120>
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:7. <https://doi.org/10.1186/s13742-015-0047-8>
- DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498. <https://doi.org/10.1038/ng.806>
- Dixon MJ, Marazita ML, Beaty TH, Murray JC (2011) Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167–178. <https://doi.org/10.1038/nrg2933>
- Gouveia MH et al (2019) Origins, admixture dynamics and homogenization of the African gene pool in the Americas. *bioRxiv*. <https://doi.org/10.1097/ede.0b013e3182125f9c>
- Grosen D et al (2011) Risk of oral clefts in twins. *Epidemiology (Cambridge, Mass)* 22:313–319. <https://doi.org/10.1101/652701>
- Hazelett DJ et al (2016) Reducing GWAS complexity. *Cell Cycle (Georgetown, Tex)* 15:22–24. <https://doi.org/10.1080/15384101.2015.1120928>
- He Z et al (2014) Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet* 94:33–46. <https://doi.org/10.1016/j.ajhg.2013.11.021>
- Karczewski KJ et al (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. <https://doi.org/10.1101/531210>
- Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS (2018) Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol* 19:14. <https://doi.org/10.1186/s13059-018-1387-3>
- Laird NM, Lange C (2008) Family-based methods for linkage and association analysis. *Adv Genet* 60:219–252. [https://doi.org/10.1016/s0065-2660\(07\)00410-5](https://doi.org/10.1016/s0065-2660(07)00410-5)
- Leslie EJ et al (2015) Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *Am J Hum Genet* 96:397–411. <https://doi.org/10.1016/j.ajhg.2015.01.004>
- Leslie EJ et al (2016) A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24.2, 17q23 and 19q13. *Hum Mol Genet* 25:2862–2872. <https://doi.org/10.1093/hmg/ddw104>
- Ludwig KU 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. <https://doi.org/10.1038/ng.2360>
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)* 26:2867–2873. <https://doi.org/10.1093/bioinformatics/btq559>
- Manolio TA et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753. <https://doi.org/10.1038/nature08494>
- McKenna A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. <https://doi.org/10.1101/gr.107524.110>
- McLaren W et al (2016) The ensembl variant effect predictor. *Genome Biol* 17:122. <https://doi.org/10.1186/s13059-016-0974-4>
- Murray T et al (2012) Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between Asians and Europeans. *Genet Epidemiol* 36:392–399. <https://doi.org/10.1002/gepi.21633>
- Naros A, Brocks A, Kluba S, Reinert S, Krimmel M (2018) Health-related quality of life in cleft lip and/or palate patients—a cross-sectional study from preschool age until adolescence. *J Craniomaxillofac Surg* 46:1758–1763. <https://doi.org/10.1016/j.jcms.2018.07.004>
- Nidey N, Wehby G (2019) Barriers to health care for children with orofacial clefts: a systematic literature review and recommendations for research priorities. *Oral Health Dent Stud* 2(1):2
- Nidey N, Moreno Uribe LM, Marazita MM, Wehby GL (2016) Psychosocial well-being of parents of children with oral clefts. *Child Care Health Dev* 42:42–50. <https://doi.org/10.1111/cch.12276>
- Poon H, Quirk C, DeZiel C, Heckerman D (2014) Literome: pubmed-scale genomic knowledge base in the cloud. *Bioinformatics (Oxford, England)* 30:2840–2842. <https://doi.org/10.1093/bioinformatics/btu383>
- Purcell S, Chang C (2019) PLINK version 1.9, cited October 2019, available at URL: <http://www.cog-genomics.org/plink/1.9/>
- Rahimov F, Jugessur A, Murray JC (2012) Genetics of nonsyndromic orofacial cleft. *Cleft Palate-craniofac J* 49:73–91. <https://doi.org/10.1597/10-178>
- Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423–449. [https://doi.org/10.1002/\(sici\)1098-2272\(1996\)13:5%3c423:Aid-gepi1%3e3.0.Co;2-3](https://doi.org/10.1002/(sici)1098-2272(1996)13:5%3c423:Aid-gepi1%3e3.0.Co;2-3)
- Schwender H, Taub MA, Beaty TH, Marazita ML, Ruczinski I (2012) Rapid testing of SNPs and gene-environment interactions in case-parent trio data based on exact analytic parameter estimation. *Biometrics* 68:766–773. <https://doi.org/10.1111/1j.1541-0420.2011.01713.x>
- Schwender H et al (2014) Detecting disease variants in case-parent trio studies using the bioconductor software package trio. *Genet Epidemiol* 38:516–522. <https://doi.org/10.1002/gepi.21836>
- Spielman RS, McGinnis RE, Ewens WJ (1994) The transmission/disequilibrium test detects cosegregation and linkage. *Am J Hum Genet* 54:559–560 (author reply 560–553)
- Taioli E, Ragin C, Robertson L, Linkov F, Thurman NE, Vieira AR (2010) Cleft lip and palate in family members of cancer survivors. *Cancer Invest* 28:958–962. <https://doi.org/10.3109/07357907.2010.483510>
- Tak YG, Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenet Chromatin* 8:57. <https://doi.org/10.1186/s13072-015-0050-4>
- Van der Auwera GA et al (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinform* 43:11.10.11–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>
- Wainschtein P et al (2019) Recovery of trait heritability from whole genome sequence data. *bioRxiv*. <https://doi.org/10.1101/588020>
- Ward LD, Kellis M (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* 44:D877–881. <https://doi.org/10.1093/nar/gkv1340>
- Wehby GL, Cassell CH (2010) The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis* 16:3–10. <https://doi.org/10.1111/j.1601-0825.2009.01588.x>
- Welsh IC, Hagge-Greenberg A, O'Brien TP (2007) A dosage-dependent role for Spry2 in growth and patterning during palate development. *Mech Dev* 124:746–761. <https://doi.org/10.1016/j.mod.2007.06.007>
- Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J (2018) High-resolution epigenomic atlas of human embryonic

craniofacial development. *Cell Rep* 23:1581–1597. <https://doi.org/10.1016/j.celrep.2018.03.129>

Zhu Y, Tazearsan C, Suh Y (2017) Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Exp Biol Med* (Maywood, NJ) 242:1325–1334. <https://doi.org/10.1177/1535370217713750>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Nandita Mukhopadhyay¹  · **Madison Bishop**² · **Michael Mortillo**³ · **Pankaj Chopra**² · **Jacqueline B. Hetmanski**⁴ · **Margaret A. Taub**⁵ · **Lina M. Moreno**⁶ · **Luz Consuelo Valencia-Ramirez**⁷ · **Claudia Restrepo**⁷ · **George L. Wehby**⁸ · **Jacqueline T. Hecht**⁹ · **Frederic Deleyiannis**¹⁰ · **Azeez Butali**¹¹ · **Seth M. Weinberg**^{1,12} · **Terri H. Beaty**⁴ · **Jeffrey C. Murray**¹³ · **Elizabeth J. Leslie**² · **Eleanor Feingold**^{1,12,14} · **Mary L. Marazita**^{1,12,15}

¹ Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Bridgeside Point Suite 500, 100 Technology Dr., Pittsburgh, PA 15219, USA

² Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA 30322, USA

³ Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

⁴ Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

⁵ Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

⁶ Department of Orthodontics, College of Dentistry, University of Iowa, Iowa City, IA 52242, USA

⁷ Fundación Clínica Noel (<http://www.clinicanoel.org.co/>), Medellín, Colombia

⁸ Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52242, USA

⁹ Department of Pediatrics, McGovern Medical School and School of Dentistry, UT Health at Houston, Houston, TX 77030, USA

¹⁰ UC Health Plastic and Reconstructive Surgery, Colorado Springs, CO 80907, USA

¹¹ Iowa Institute of Oral Health Research, College of Dentistry, University of Iowa, Iowa City, IA 52242, USA

¹² Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15219, USA

¹³ Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, IA 52242, USA

¹⁴ Department of Biostatistics Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA

¹⁵ Department of Psychiatry, School of Medicine and Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA 15260, USA