bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Nandita Mukhopadhyay[1], 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. Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, 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/), Medellin, 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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, Iowa, 52242, USA
14. Department of Biostatistics Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA
15. Clinical and Translational Science Institute, and Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15262, USA

§ Co-Senior Authors

* CORRESPONDING AUTHOR:
Mary L. Marazita, PhD.
Center for Craniofacial and Dental Genetics
Bridgeside Point Suite 500
100 Technology Dr.
Pittsburgh, PA 15219
phone: 412-648-8380
FAX: 412-648-8387
Email: marazita@pitt.edu

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Orofacial clefts (OFCs) are one of the most common birth defects worldwide and create a significant health burden. The majority of OFCs are non-syndromic, and the genetic component has been 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. Part of the Gabriella Miller Kids First Pediatric Research Program, this is the first large-scale WGS study of OFC in parent-offspring trios. WGS provides deeper and more specific genetic data than currently available using imputation on single nucleotide polymorphic (SNP) marker panels. Here, association analysis of genome-wide single nucleotide variants (SNV) and short insertions and deletions (indels) identified a new locus on chromosome 21 in Colombian families, within a region known to be expressed during craniofacial development. This study reinforces the ancestry differences seen in the genetic etiology of OFCs, and the need for larger samples when for studying OFCs and other birth defects in admixed populations.

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 (1, 2). Surgical treatment along with ongoing orthodontia, speech and other therapies, are 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 (3, 4). Furthermore, there are disparities in access to such therapies for OFCs (5), similar to other malformations with complex medical and surgical needs. Some studies have suggested a reduced quality of life for individuals with OFCs (6), while other studies have identified higher

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

risk to certain types of cancers (7-9). 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.ncbi.nlm.nih.gov/omim, search criterion=(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 non-syndromic, i.e. they occur as isolated anomalies with no other apparent cognitive or structural abnormalities (1). 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) (10), but as for most other complex human traits (11-13), very few putative functional variants for non-syndromic OFCs have been identified from genome-wide association studies (GWASs)(14). In particular, the high heritability for OFC, estimated at 90% by a twin study in a Danish sample (15) cannot be explained by all identified common variants significantly associated with OFC, sometimes referred to as the "missing heritability" problem (16). Additional approaches will be necessary to expand our understanding of genetic variation in nonsyndromic 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 (17).

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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (7, 8), 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 cleft lip with or without cleft palate (CL/P).

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 the MAF cutoff). A comparison of the p-values between allelic TDT (aTDT) and genotypic TDT (gTDT) showed high concordance (see section "Comparison between aTDT and gTDT" and supplementary figure S1 Figure 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 1 and 2 show the most significant results in the **European** (Table 1) and **Colombian** trios (Table 2). Several SNPs gave genome-wide significant association p-values in the stratified

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

aTDT analysis of **European** (Table 1 and Figure 1 top panel) and **Colombian** trios (Table 2 and Figure 1 middle panel), and a single SNP achieved genome-wide significance in the **Combined** sample (Figure 1 bottom panel). In the **European** sample, 17 significant associations are observed across multiple chromosomes (Table 1). In the **Colombian** sample, four significant associations are observed for markers on chromosomes 6, 8, 19 and 21. After close examination of the 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 (18). 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 1 lists all **European** peaks and contains the smallest association p-values with their

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

143 corresponding estimated odds ratios (OR) observed in the **Colombian** and **Combined** aTDTs
144 within 500 KB on either side of each **European** peak SNP (Table 1 columns 4-7). Since allele
145 frequencies for specific SNPs may differ between the two samples, this provides a region-level
146 view of replication across the samples. Similarly, Table 2 lists the **Colombian** peaks, along with
147 the minimum association p-values and corresponding odds ratios observed in the **European** and
148 **Combined** aTDTs within 500 KB on either side of each **Colombian** peak. As seen in Tables 1
149 and 2, **European** and **Colombian** trios differ considerably with respect to the genomic regions
150 that show significant association to CL/P.

151 Table 1. 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**	**p-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**	
			p-value (OR)	**RS number (bp position)**	**p-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)

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

Table 2. 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**	**p 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**	
			p value (OR)	**RS number (bp position)**	**p 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)

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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

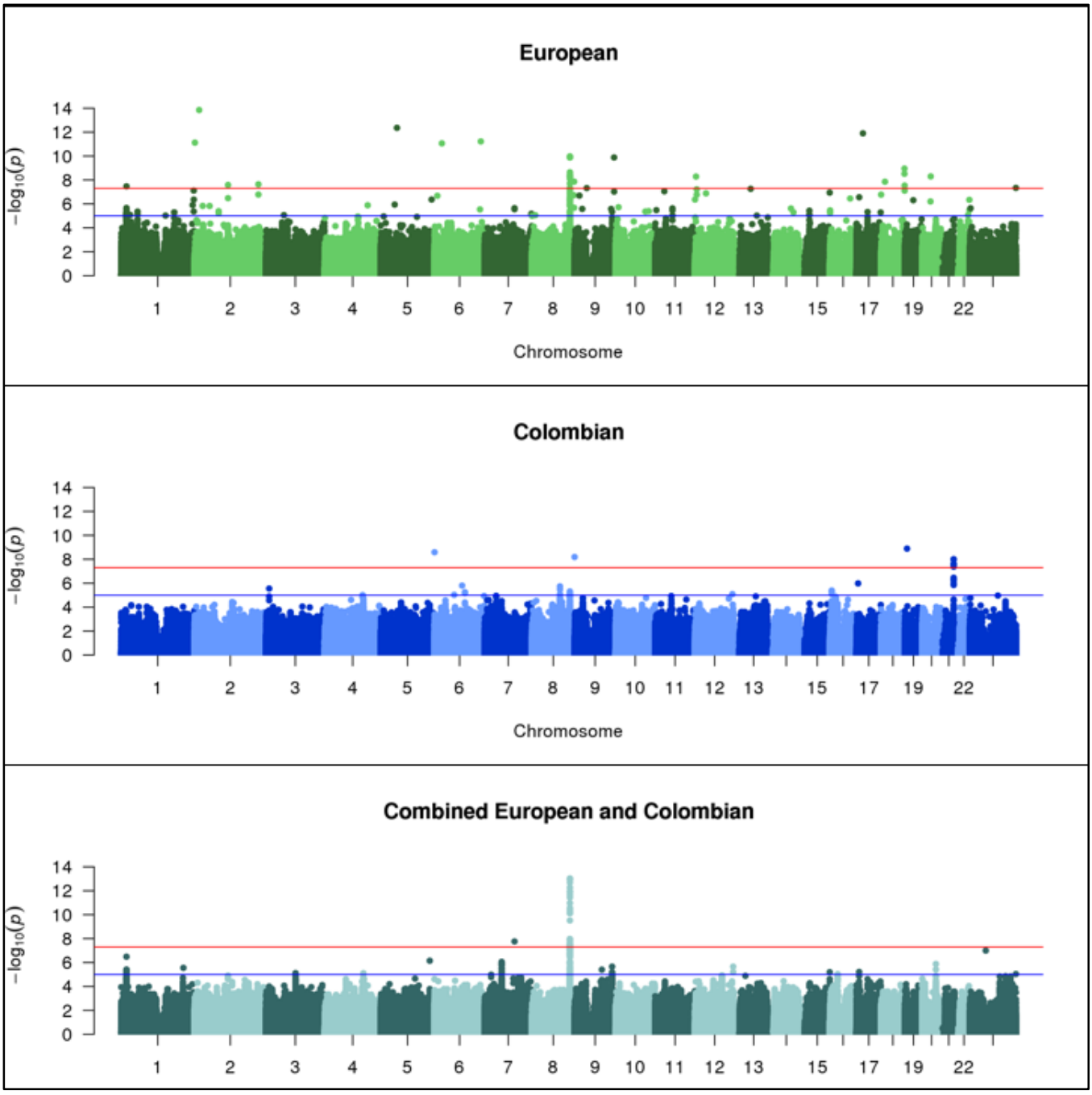

Figure 1. Manhattan plots of **European (315 trios)**, **Colombian (265 trios)** and **Combined (580 trios)** allelic TDTs

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 others (19-21). The 1p36.13 peak is located 23kb upstream of the transcription start site of the *PAX7* gene. These associations were

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (22).

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. This SNP 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 52kb centromeric to the *NTN1* gene (p=2.77e-07, odds ratio=0.29, CI=[0.18−0.48]). None of the other previously reported OFC variants 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. Supplement S2 Table 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 30kb interval on chromosome 21q22.3 (Figure 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 two-fold increases in the transmission of the risk alleles from parents to the proband offspring. The smallest p-value was observed at rs2839575 (p=9.75e-09, odds ratio=2.48, 95% CI = [1.81−3.45]).

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

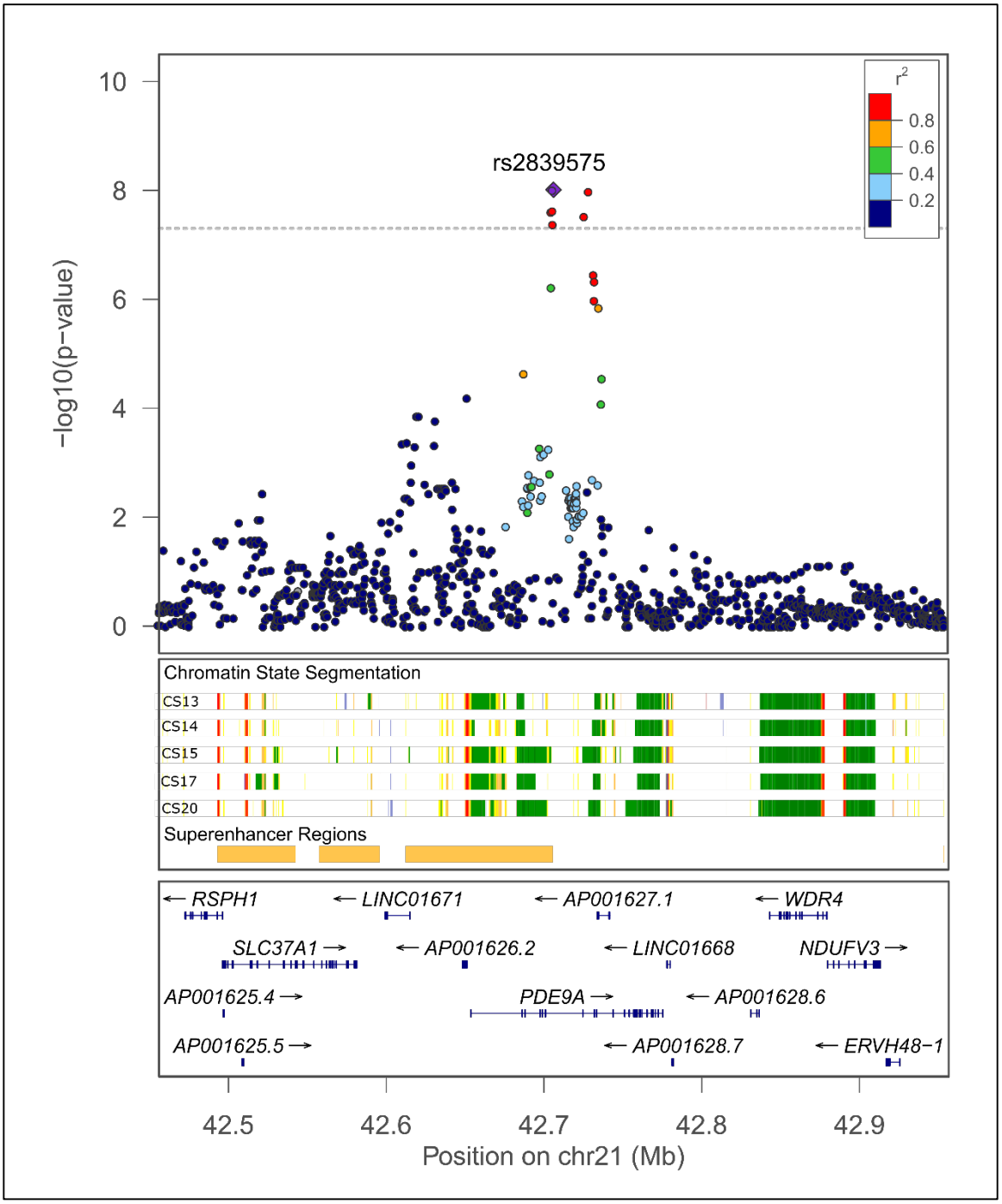

Figure 2. $Log_{10}(p)$ for SNPs and indels in chromosome 21q22.3 peak region.

GWAS of a Latino sample from a previous study, the POFC Multiethnic study (24), reported suggestive association at this genomic region (see Figure 1 in (23)). That Latino sample included diverse Hispanic groups from the US, Guatemala, Argentina, and Colombia, and all of the current WGS

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

189 **Colombia** trios were part of the POFC Multiethnic study. However, the POFC Multiethnic study had
190 129 additional Colombian trios. In that study, the GWASs of Asian and European samples did not show
191 association in this region, and nor did the combined GWAS of all the POFC Multiethnic study samples.
192 The fact that the current WGS case-parent trio study yielded a genome-wide significant association with
193 a smaller sample size suggests this association might be unique to **Colombians**. We explored the
194 validity and implications of this observation through a number of analyses, as described below.
195 We first examined the aTDT p-values for our **Colombian** WGS trios using their SNP array data
196 from the POFC Multiethnic study. The p-values in this region were nearly identical to those observed in
197 our WGS association, confirming the association we observed here was not an artifact of sequencing.
198 We next investigated whether population substructure within the Colombian parents could have
199 caused the observed association in the WGS data by examining the ancestry principal components (PCs)
200 as well as results of quantitative association between PCA eigenvalues to variants within the peak region
201 (see Methods for details). PCA showed no evidence of population substructure (supplementary figure S1
202 Figure 2A), and no association was observed between the eigenvalues and variants in the chromosome
203 21q22.3 region (supplementary figure S1 Figure 2B). A positive association between eigenvalues and
204 variants would have indicated that the observed association with CL/P is in reality due to population
205 substructure, therefore, this association did not appear to be an artifact of population admixture.
206 We verified that this region does not show evidence of association in other Latin Americans, by
207 reanalyzing imputed genotype data from the previously published POFC Multiethnic GWAS study (23).
208 P-values from the aTDT of independent trios and the corresponding odds ratios at rs2839575 in each
209 Latino subpopulation were considerably different from those observed in the Colombian subjects
210 (Figure 3A, forest plot). In fact, in contrast to the Colombians, none of the other Latino populations
211 (except Colombia) showed a significant association at rs2839575. Moreover, the combined set of non-

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Figure 3B, regional p-value plot and supplementary table S3 Table). 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 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 Figure 3).

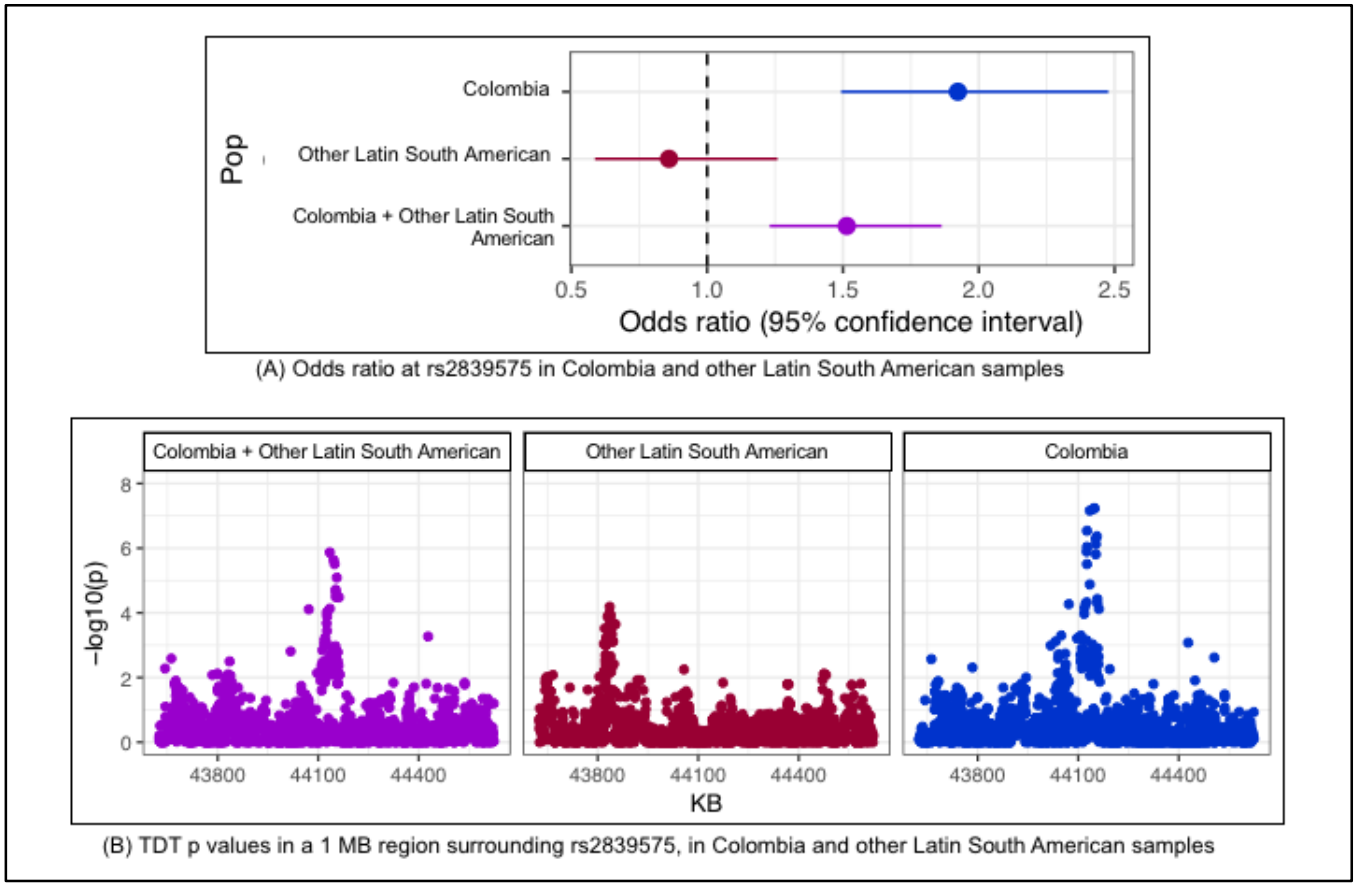

Figure 3. Estimated odds ratios (with 95%CI) and -log10(p-values) from the aTDT in Colombia and other Latino samples

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 are 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 (24). 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 Figure 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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

248 associations were all intronic variants within the *PDE9A* gene; however, all had moderate odds ratios
249 around 2.0. In this region, there were 37 SNPs with minor allele frequencies near or below 1% in the
250 **Colombian** trios and estimated OR > 5 (supplementary table S4 Table), including mainly intronic and a
251 few intergenic SNVs (28 intronic, 8 intergenic). The exception was one non-synonymous SNV,
252 rs138007679 in the *RSPH1* gene (aTDT odds ratio 8, 95% CI = [1.001−63.96]), which produces an
253 amino acid change (A>C, leucine to tryptophan according to ClinVar). Alone, this variant does not
254 clearly implicate *RSPH1* over other genes in the region, so we performed a rare variant TDT on all non-
255 synonymous variants within the 13 genes falling in this region. None of the individual genes achieved
256 the nominal significance (supplementary table S5 Table), so this result remained inconclusive. We also
257 carried out rare variant TDTs of intronic and intergenic variants with similar results, finding only
258 nominally significant associations attributable to intergenic, low frequency variants (MAFs ranging
259 between 0.5% and 1%) .

260 In the absence of any clearly pathogenic variant or gene based on combined effects of rare
261 variants, we examined regulatory elements and protein-protein interaction pathways in this region with
262 respect to craniofacial development. All associated variants below a suggestive level of significance
263 ($p<1.0e-05$) were located within the *PDE9A* gene, which does not have any known role in controlling
264 risk to OFCs. However, the *PDE9A* gene overlaps a super-enhancer region for craniofacial development
265 identified from histone profiling in early human craniofacial development (25). Multiple genes in the
266 region, including *PDE9A*, appear to be actively transcribed during human craniofacial development
267 (Figure 2). Another gene of interest is *UBASH3A*, located ~220kb centromeric to this peak signal. The
268 UBASH3A protein was previously shown to physically associate with *SPRY2* via a yeast two-hybrid
269 assay (33). SPRY2 has been reported by GWASs of OFC and shown required for palatogenesis in mice
270 (26); whether UBASH3A is also expressed in craniofacial structures has not yet been determined.

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (27). 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 to 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 (23). 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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Research Design and 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.

This study included case-parent trios consisting of affected offspring and their parents (see Table 3). Most of the **European** parents and all **Colombian** parents are unaffected for CL/P (see breakdown of trios in Table 3). 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 3 shows the counts of GMKF trios sequenced for the present study, by their country of origin.

Sample	Total Trios	Trios with no affected parents	Trios with 1 affected parent	Trios with 2 affected parents
European	315	280	32	3

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Table 3. Counts of CL/P trios by recruitment site and cleft type (add?)

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 30X 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 (28-30) at the Broad Institute (https://software.broadinstitute.org/gatk/best-practices/workflow). The alignment and joint genotyping workflow used to 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 was 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

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 in the supplement, S1 section "Kids First DRC Genomics Harmonization Pipeline Description".

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 (Ts/Tv > 3 standard deviations from mean Ts/Tv 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 (23) 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 1-10,000 BP in length) with a genotyping rate of at least 90% were included in our analyses. A total of 5,374,579

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

358 variants were analyzed in the **European** trios, and 4,905,638 in **Colombian** trios. Of these, 4,220,712
359 variants were analyzed for the **Combined** trios.

360 Genome-wide wide association testing of SNPs and indels

361 Genome-wide association was conducted using two versions (allelic and genotypic) of the
362 transmission disequilibrium test (TDT), for each polymorphic variant. The PLINK software (31, 32) was
363 used to run the standard genome-wide allelic TDT (aTDT), which does not consider the parents' cleft
364 status. We also ran genotypic TDT (gTDT) (33) on the trios, and compared the association p-values to
365 those from the aTDT. Effect sizes are not directly comparable between the two methods. The aTDT
366 compares the transmission of a target allele to the affected child from heterozygous parents (34), and is
367 based on McNemar's chi-square statistic. Because only heterozygous parents can contribute to this
368 statistic, statistical power is greatly influenced by minor allele frequency (MAF) and one population
369 may have considerably more or less power at any given SNP when MAF varies across populations. The
370 gTDT compares the observed genotype in the child to "pseudo-controls" representing other genotypes
371 possible from the parental mating type. Schwender et al. (35, 36) demonstrated an efficient method for
372 computing this gTDT statistic. Because either TDT represents a test of strict Mendelian inheritance of
373 the marker (despite sampling case-parent trios through the affected proband), this test is robust to
374 spurious associations arising from population stratification and can provide greater power for rare
375 phenotypes (37). The null hypothesis of either TDT is the complete absence of either linkage between
376 the marker and an unobserved causal gene or linkage disequilibrium (LD) between the marker and an
377 unobserved causal gene. Rejection of this composite null hypothesis implies the presence of both
378 linkage and LD. The TDT is most appropriate for our study, given our participants originate from
379 diverse populations, and the Colombians in particular are known to reflect varying degrees of admixture
380 of African, Hispanic, Native American and European genes.

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Three genome-wide TDT analyses were run: separately in European and Colombian trios and then in all trios combined. Significance 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 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 (38). PLINK (31, 32) 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 1Mb 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 (39), Variant Effect Predictor (40), and

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HaploReg (41). 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 (42).

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 (43) using all variants within a selected region. The RV-TDT software (44) was then run on phased haplotypes for exonic, non-synonymous SNVs in genes with a minimum of 4 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 receiving 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 up into subsets based on gene locations in this region, and analyzed using a procedure similar to the exonic, non-synonymous SNPs.

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 McDonell Genome Institute was

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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). 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). 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, R03-DE026469, R01-DE012472, U01-DD000295, R01-DE014581, R01-DE011931, R37-DE008559, R21-DE016930, and R01-DE015667, R03-DE027193, R00-DE025060].

Many thanks to the participating families and study teams worldwide without whom these studies would not be possible. Additional thanks to Andrew Lidral, Mauricio Arcos-Burgos, and Andrew Czeizel.

References

1 Dixon, M.J., Marazita, M.L., Beaty, T.H. and Murray, J.C. (2011) Cleft lip and palate: understanding genetic and environmental influences. *Nature reviews. Genetics*, **12**, 167-178.

2 Rahimov, F., Jugessur, A. and Murray, J.C. (2012) Genetics of nonsyndromic orofacial clefts. *The Cleft palate-craniofacial journal : official publication of the American Cleft Palate-Craniofacial Association*, **49**, 73-91.

3 Nidey, N., Moreno Uribe, L.M., Marazita, M.M. and Wehby, G.L. (2016) Psychosocial well-being of parents of children with oral clefts. *Child: care, health and development*, **42**, 42-50.

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Wehby, G.L. and Cassell, C.H. (2010) The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral diseases*, **16**, 3-10.

5 Nidey, N. and Wehby, G. (2019) Barriers to Health Care for Children with Orofacial Clefts: A Systematic Literature Review and Recommendations for Research Priorities. *Oral Health and Dental Studies*, **2(1):2**.

6 Naros, A., Brocks, A., Kluba, S., Reinert, S. and Krimmel, M. (2018) Health-related quality of life in cleft lip and/or palate patients - A cross-sectional study from preschool age until adolescence. *Journal of cranio-maxillo-facial surgery : official publication of the European Association for Cranio-Maxillo-Facial Surgery*, **46**, 1758-1763.

7 Bille, C., Winther, J.F., Bautz, A., Murray, J.C., Olsen, J. and Christensen, K. (2005) Cancer risk in persons with oral cleft--a population-based study of 8,093 cases. *American journal of epidemiology*, **161**, 1047-1055.

8 Bui, A.H., Ayub, A., Ahmed, M.K., Taioli, E. and Taub, P.J. (2018) Association Between Cleft Lip and/or Cleft Palate and Family History of Cancer: A Case-Control Study. *Annals of plastic surgery*, **80**, S178-s181.

9 Taioli, E., Ragin, C., Robertson, L., Linkov, F., Thurman, N.E. and Vieira, A.R. (2010) Cleft lip and palate in family members of cancer survivors. *Cancer investigation*, **28**, 958-962.

10 Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, **47**, D1005-d1012.

11 Hazelett, D.J., Conti, D.V., Han, Y., Al Olama, A.A., Easton, D., Eeles, R.A., Kote-Jarai, Z., Haiman, C.A. and Coetzee, G.A. (2016) Reducing GWAS Complexity. *Cell cycle (Georgetown, Tex.)*, **15**, 22-24.

12 Tak, Y.G. and Farnham, P.J. (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. *Epigenetics & chromatin*, **8**, 57.

13 Zhu, Y., Tazearslan, C. and Suh, Y. (2017) Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Experimental biology and medicine (Maywood, N.J.)*, **242**, 1325-1334.

14 Beaty, T.H., Marazita, M.L. and Leslie, E.J. (2016) Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Research*, **5**, 2800.

15 Grosen, D., Bille, C., Petersen, I., Skytthe, A., Hjelmborg, J., Pedersen, J.K., Murray, J.C. and Christensen, K. (2011) Risk of oral clefts in twins. *Epidemiology (Cambridge, Mass.)*, **22**, 313-319.

16 Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

17 Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M. *et al.* (2019) Recovery of trait heritability from whole genome sequence data. in press., 588020.

18 Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. *et al.* (2019) Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. in press., 531210.

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19 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. *Nature genetics*, **44**, 968-971.

20 Birnbaum, S., Ludwig, K.U., Reutter, H., Herms, S., de Assis, N.A., Diaz-Lacava, A., Barth, S., Lauster, C., Schmidt, G., Scheer, M. *et al.* (2009) IRF6 gene variants in Central European patients with non-syndromic cleft lip with or without cleft palate. *European journal of oral sciences*, **117**, 766-769.

21 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. *Nature genetics*, **42**, 525-529.

22 Leslie, E.J., Taub, M.A., Liu, H., Steinberg, K.M., Koboldt, D.C., Zhang, Q., Carlson, J.C., Hetmanski, J.B., Wang, H., Larson, D.E. *et al.* (2015) Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *American journal of human genetics*, **96**, 397-411.

23 Leslie, E.J., Carlson, J.C., Shaffer, J.R., Feingold, E., Wehby, G., Laurie, C.A., Jain, D., Laurie, C.C., Doheny, K.F., McHenry, T. *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. *Human molecular genetics*, **25**, 2862-2872.

24 Gouveia, M.H., Borda, V., Leal, T.P., Moreira, R.G., Bergen, A.W., Aquino, M.M., Araujo, G.S., Araujo, N.M., Kehdy, F.S.G., Liboredo, R. *et al.* (2019) Origins, admixture dynamics and homogenization of the African gene pool in the Americas. in press., 652701.

25 Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P. and Cotney, J. (2018) High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. *Cell reports*, **23**, 1581-1597.

26 Welsh, I.C., Hagge-Greenberg, A. and O'Brien, T.P. (2007) A dosage-dependent role for Spry2 in growth and patterning during palate development. *Mech Dev*, **124**, 746-761.

27 Murray, T., Taub, M.A., Ruczinski, I., Scott, A.F., Hetmanski, J.B., Schwender, H., Patel, P., Zhang, T.X., Munger, R.G., Wilcox, A.J. *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. *Genetic epidemiology*, **36**, 392-399.

28 DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491-498.

29 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297-1303.

30 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J. *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics*, **43**, 11.10.11-33.

31 Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

32 Purcell, S. and Chang, C., in press.

33 Schaid, D.J. (1996) General score tests for associations of genetic markers with disease using cases and their parents. *Genetic epidemiology*, **13**, 423-449.

34 Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1994) The transmission/disequilibrium test detects cosegregation and linkage. *American journal of human genetics*, **54**, 559-560; author reply 560-553.

bioRxiv preprint doi: https://doi.org/10.1101/743526; this version posted August 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

543 35 Schwender, H., Li, Q., Neumann, C., Taub, M.A., Younkin, S.G., Berger, P., Scharpf, R.B.,
544 Beaty, T.H. and Ruczinski, I. (2014) Detecting disease variants in case-parent trio studies using the
545 bioconductor software package trio. *Genetic epidemiology*, **38**, 516-522.
546 36 Schwender, H., Taub, M.A., Beaty, T.H., Marazita, M.L. and Ruczinski, I. (2012) Rapid testing
547 of SNPs and gene-environment interactions in case-parent trio data based on exact analytic parameter
548 estimation. *Biometrics*, **68**, 766-773.
549 37 Laird, N.M. and Lange, C. (2008) Family-based methods for linkage and association analysis.
550 *Advances in genetics*, **60**, 219-252.
551 38 Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. and Chen, W.M. (2010)
552 Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*,
553 **26**, 2867-2873.
554 39 Kotlar, A.V., Trevino, C.E., Zwick, M.E., Cutler, D.J. and Wingo, T.S. (2018) Bystro: rapid
555 online variant annotation and natural-language filtering at whole-genome scale. *Genome biology*, **19**, 14.
556 40 McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and
557 Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome biology*, **17**, 122.
558 41 Ward, L.D. and Kellis, M. (2016) HaploReg v4: systematic mining of putative causal variants,
559 cell types, regulators and target genes for human complex traits and disease. *Nucleic acids research*, **44**,
560 D877-881.
561 42 Poon, H., Quirk, C., DeZiel, C. and Heckerman, D. (2014) Literome: PubMed-scale genomic
562 knowledge base in the cloud. *Bioinformatics (Oxford, England)*, **30**, 2840-2842.
563 43 Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-
564 data inference for whole-genome association studies by use of localized haplotype clustering. *American*
565 *journal of human genetics*, **81**, 1084-1097.
566 44 He, Z., O'Roak, B.J., Smith, J.D., Wang, G., Hooker, S., Santos-Cortez, R.L., Li, B., Kan, M.,
567 Krumm, N., Nickerson, D.A. *et al.* (2014) Rare-variant extensions of the transmission disequilibrium
568 test: application to autism exome sequence data. *American journal of human genetics*, **94**, 33-46.
569