ASSOCIATION STUDIES ARTICLE

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

Elizabeth J. Leslie¹, Jenna C. Carlson¹,², John R. Shaffer¹,³, Eleanor Feingold¹,²,³, George Wehby⁴, Cecelia A. Laurie⁵, Deepti Jain⁵, Cathy C. Laurie⁵, Kimberly F. Doheny⁶, Toby McHenry¹, Judith Resick¹, Carla Sanchez¹, Jennifer Jacobs¹, Beth Emanuele¹, Alexandre R. Vieira¹,³, Katherine Neiwanger¹, Andrew C. Lidral⁷, Luz Consuelo Valencia-Ramirez¹¹, Ana Maria Lopez-Palacio¹², Dora Rivera Valencia¹³, Mauricio Arcos-Burgos¹⁴, Andrew E. Czeizel¹⁵, L. Leigh Field¹⁶, Carmencita D. Padilla¹⁷,¹⁸, Eva Maria, C. Cutiongco-de la Paz¹⁷,¹⁸, Frederic Deleyiannis¹⁹, Kaare Christensen²⁰, Ronald G. Munger²¹, Rolv T. Lie²², Allen Wilcox²³, Paul A. Romitti⁸, Eduardo E. Castilla²⁴,²⁵,²⁶, Juan C. Mereb²⁸, Fernando A. Poletta²⁴,²⁵,²⁶, Iêda M. Orioli²⁶,²⁷, Flavia M. Carvalho²⁵,²⁶, Jacqueline T. Hecht²⁹, Susan H. Blanton³⁰, Carmen J. Buxó³¹, Azeez Butali⁹, Peter A. Mossey³², Wasiu L. Adeyemo³³, Olutayo James³³, Ramat O. Braimah³⁴, Babatunde S. Aregebosa³⁴, Mekonen A. Eshete³⁵, Fikre Abate³⁵, Mine Koruyucu³⁶, Figen Seymen³⁶, Lian Ma³⁷, Javier Enríquez de Salamanca³⁸, Seth M. Weinberg¹, Lina Moreno⁷, Jeffrey C. Murray¹⁰ and Mary L. Marazita¹,³,³⁹,*

¹Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA 15219, USA, ²Department of Biostatistics and ³Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA, ⁴Department of Biostatistics and ⁵Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52246, USA, ⁶Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ⁷Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD 21224, USA, ⁸Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52246, USA, ⁹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁰Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52246, USA, ¹¹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹²Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹³Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁴Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁵Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁶Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁷Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁸Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ¹⁹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁰Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²¹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²²Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²³Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁴Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁵Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁶Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁷Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁸Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ²⁹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁰Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³¹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³²Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³³Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁴Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁵Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁶Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁷Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁸Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, ³⁹Department of Biostatistics, Genetic Coordinating Center, University of Washington, Seattle, WA 98195, USA, *Corresponding author, E-mail: marazita@upmc.edu

We dedicate this paper to the memory of our colleagues Andrew Czeizel and Dr. Juan C. Mereb; they will be missed.

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Received: January 17, 2016. Revised: March 4, 2016. Accepted: March 24, 2016

© The Author 2016. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com
Abstract

Orofacial clefts (OFCs), which include non-syndromic cleft lip with or without cleft palate (CL/P), are among the most common birth defects in humans, affecting approximately 1 in 700 newborns. CL/P is phenotypically heterogeneous and has a complex etiology caused by genetic and environmental factors. Previous genome-wide association studies (GWASs) have identified at least 15 risk loci for CL/P. As these loci do not account for all of the genetic variance of CL/P, we hypothesized the existence of additional risk loci. We conducted a multiethnic GWAS in 6480 participants (823 unrelated cases, 1700 unrelated controls and 1319 case-parent trios) with European, Asian, African and Central and South American ancestry. Our GWAS revealed novel associations on 2p24 near FAM49A, a gene of unknown function (P = 4.22 × 10⁻⁸), and 19q13 near RHPN2, a gene involved in organizing the actin cytoskeleton (P = 4.17 × 10⁻⁹). Other regions reaching genome-wide significance were 1p36 (PAX7), 1p22 (ARHGAP29), 1q32 (IRF6), 8q24 and 17p13 (NTN1), all reported in previous GWASs. Stratification by ancestry group revealed a novel association with a region on 17q23 (P = 2.92 × 10⁻⁵) among individuals with European ancestry. This region included several promising candidates including TANC2, an oncogene required for development, and DCAF7, a scaffolding protein required for craniofacial development. In the Central and South American ancestry group, significant associations with loci previously identified in Asian or European ancestry groups reflected their admixed ancestry. In summary, we have identified novel CL/P risk loci and suggest new genes involved in craniofacial development, confirming the highly heterogeneous etiology of OFCs.
Introduction

Orofacial clefts (OFCs) are a heterogeneous group of craniofacial malformations that comprise a significant fraction of all human birth defects, occurring in approximately 1 in 700 individuals worldwide (1). The prevalence of OFCs varies widely across populations as Native American and Asian populations have the highest rates (~1/500 live births), whereas African and African-American populations have the lowest rates (~1/25000 live births). European-derived populations have intermediate rates at approximately 1/1000 live births. Populations from Central and South America also have intermediate prevalence rates that vary regionally, possibly reflecting their mixed ancestries.

Non-syndromic OFCs, which occur in the absence of other cognitive or structural abnormalities, have a complex etiology that reflects the combined actions of multiple genetic and environmental risk factors. The focus of much of the genetics research has been on the common forms of OFCs: cleft lip alone (CL), CL with palate (CLP) and cleft palate alone (CP) (2). As true for many complex traits, substantial progress in gene identification has occurred for OFCs, owing to multiple successful genome-wide association studies (GWASs). To date, there have been six OFC GWASs (3–8), a meta-analysis (9) and several replication studies (10–12). A major finding from this work is that OFCs have significant genetic heterogeneity; collectively, these studies have identified at least 15 genetic loci with compelling statistical and biological support. However, these loci account only for a modest portion of the genetic variance of CLP, suggesting that additional genetic risk factors are likely involved.

In addition, some of these loci appear to have far stronger effect sizes in some populations (3,9), which may contribute to the variability in prevalence rates. For example, the 8q24 locus shows strong evidence of association in populations with European ancestry, whereas the 1q32 (IRF6) and 20q12 (MABF) loci show stronger effects among Asian populations (3,13). However, the majority of GWASs to date have focused on populations with European ancestry; only two included Asian populations (3,8,10), and none focused on Hispanic/Latino or African populations. In order to identify additional risk loci in diverse and understudied populations, we conducted a large multietnic GWAS comprising unrelated cases, unrelated controls and trios of European, Asian and African ancestry, as well as admixed populations from Central and South America.

Results

We performed tests of association in the two subsets of our multi-ethnic sample. The first was a set of 1319 case–parent trios in which the probands had CL/P, and the second was an independent case–control set consisting of 823 unrelated cases with CL/P and 1700 unrelated controls (Supplementary Material, Fig. S3). The 293,633 genotyped single nucleotide polymorphisms (SNPs) and 33,669,354 imputed SNPs were analyzed separately in each subset and combined by meta-analysis. Here, we focus on the results of the meta-analysis, but we note that similar results were obtained in both the case–control and case–parent trio analyses, indicating that the meta-analysis results are not driven by only one of the analyses. In the combined multiethnic sample, 316 (imputed) and 113 (genotyped) SNPs from seven loci were associated with CL/P at genome-wide significance (Fig. 1A, Table 1). Five of these loci were previously reported in GWASs of CL/P (1p36, near PAX7, 1p22 near ARHGAP29, 1q32 near IRF6, 8q24 at a gene desert and 17p13.1 near NTN1). Associations for two novel loci were observed on chromosome 2q24.2 near FAM49A (lead SNP rs7552, P = 4.22 × 10⁻⁷; Fig. 2A) and 19q13.11 near RHPN2 (lead SNP rs73039426, P = 2.92 × 10⁻⁸, Fig. 2B).

Seven additional regions approached genome-wide significance with P-values < 5 × 10⁻⁷. All genotyped SNPs with P-values < 1 × 10⁻⁵ are available in Supplementary Material, Tables S4–S7. One of these was a previously reported locus on 8q21 near DCAF4L2 and MMP16. Four novel loci had multiple SNPs approaching genome-wide significance: 5p12 near FGF10 (lead SNP rs59569344, P = 4.75 × 10⁻⁹), 8q22 in VPS13B (lead SNP rs1788160, P = 9.52 × 10⁻⁹), 10q26.3 in MGMT (lead SNP rs7922405, P = 4.60 × 10⁻⁷) and 15q24.1 in ARID3B (lead SNP rs28689146, P = 5.4 × 10⁻⁷). Regional plots for these four novel loci are in Supplementary Material, Figure S4. At the two remaining novel loci (2p22 and 9q21.3), single SNPs approached genome-wide significance.

To explore population-specific associations, we stratified our sample into genetically defined continental ancestry groups and used the same analysis approach (i.e., meta-analysis of results from the trio and case–control subsets) in European, Asian and Central/South American groups. Stratified analysis was not performed separately in the African group due to small sample size. Manhattan plots for case–control and transmission disequilibrium test (TDT) analyses in the subpopulations are presented in Supplementary Material, Figures S5–S7.

In the European group, two loci on 8q24 and 17q23 showed genome-wide significant associations (Fig. 1B, Supplementary Material, Table S5). The most significant SNP at the 8q24 gene desert was rs72728734 (P = 7.33 × 10⁻¹⁵). At the novel 17q23 locus, the most significant SNP was rs1588366 (P = 1.41 × 10⁻⁶) near TANC2 (Fig. 2C). Three loci approached genome-wide significance: 1p36 (PAX7), 17p13.1 (NTN1) and a novel locus on 6p21 at the HLA-B locus (lead SNP rs79416102, P = 2.92 × 10⁻⁷).

In the Asian group, no significant associations were observed (Fig. 1C). However, the previously reported region on 1q32 near IRF6 (lead SNP rs1044516; P = 1.01 × 10⁻⁶) and several SNPs on 17p13 (lead SNP rs12451139; P = 4.20 × 10⁻⁷) showed ‘suggestive’ evidence of association, represented by multiple SNPs with P-values in the order of magnitude range 10⁻⁷ to 10⁻⁶ (Supplementary Material, Table S6). SNPs on 1q32 and 17p13 have shown association with CL/P in Asian populations in prior studies and were observed in our multi-ethnic analysis. However, the most significant SNPs on chromosome 17 in the present study are located nearly 800 kb centromeric to the previously reported 17p13 region overlapping NTN1 and may represent a novel region unrelated to NTN1.

In the Central and South American group, three loci showed genome-wide significant evidence of association (1q32, 8q24 and 17p13) (Fig. 1D, Supplementary Material, Table S7). Five loci approached genome-wide significance on chromosomes 1p22, 6p12, 8p23, 16p13 and 21q22.3 (Table 1). We further scrutinized the peaks with suggestive evidence of association (P < 1.0 × 10⁻⁵), among which were a combination of signals originally observed in both European and Asian groups including 8q21 near MMP16 (lead SNP rs111608481, P = 3.43 × 10⁻⁶), and 17q22 near NOG (lead SNP rs227727, P = 5.73 × 10⁻⁷). The combination of signals originally observed in both European and Asian groups likely reflects the admixture and population history of the region.

We performed conditional analyses on each of the genome-wide significant peaks within the case–control subset to determine whether these peaks represent multiple independent association signals. We identified two independent signals at the 8q24 locus in Europeans (Fig. 3). One signal was observed in
Figure 1. Manhattan plots of -log10(P-values) from meta-analysis of case–control and TDT results in a multiethnic CL/P cohort. Results are shown for the combined multiethnic sample (A) and groups with (B) European ancestry, (C) Asian ancestry and (D) Central and South American ancestry. Colored lines denote suggestive (blue) and genome-wide (red) thresholds for significance. The genomic inflation factors, $\lambda$, are 1.088, 1.01, 0.979 and 1.045, for the multiethnic, European, Asian and Central/South American scans, respectively, indicating minor inflation in p-values consistent with the observation of multiple strongly associated loci.
previous studies (3–5); our lead SNP, rs72728734 (P = 6.60 × 10⁻⁶) was in high linkage disequilibrium (LD) with rs987525 (r² = 0.81), the lead SNP identified previously. The second emerged after conditioning on our lead SNP (rs184216519, P = 9.57 × 10⁻⁷). We similarly observed two independent signals in the multi-ethnic meta-analysis but did not find evidence of two signals at this locus in the Central/South American analyses (data not shown).

Conditioning on lead SNPs at 1q32 (IRF6) and 17p13 (NTN1) in the multiethnic analysis (Supplementary Material, Fig. S8) and 1q32 in the Central/South American analysis (Supplementary Material, Fig. S9) suggested the presence of multiple independent signals.

**Table 1.** Summary of GWAS meta-analysis results for loci with P < 5 × 10⁻⁷

<table>
<thead>
<tr>
<th>GWAS analysis population</th>
<th>Locus</th>
<th>SNP</th>
<th>Risk allele</th>
<th>TDT OR</th>
<th>CC OR</th>
<th>Meta OR</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-ethnic</td>
<td>1p36.13</td>
<td>rs4920524</td>
<td>G</td>
<td>1.33</td>
<td>1.34</td>
<td>1.34</td>
<td>1.21-1.47</td>
<td>3.72E⁻⁰⁹</td>
</tr>
<tr>
<td></td>
<td>1p22.1</td>
<td>rs4147882</td>
<td>T</td>
<td>1.49</td>
<td>1.33</td>
<td>1.42</td>
<td>1.28-1.58</td>
<td>5.59E⁻¹¹</td>
</tr>
<tr>
<td></td>
<td>1q32</td>
<td>rs11119345</td>
<td>G</td>
<td>1.78</td>
<td>1.85</td>
<td>1.81</td>
<td>1.57-2.07</td>
<td>2.52E⁻¹⁷</td>
</tr>
<tr>
<td></td>
<td>2p22</td>
<td>rs372148350</td>
<td>AT</td>
<td>1.37</td>
<td>1.80</td>
<td>1.49</td>
<td>1.28-1.74</td>
<td>2.50E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>2p24.2</td>
<td>rs7552</td>
<td>G</td>
<td>1.36</td>
<td>1.18</td>
<td>1.28</td>
<td>1.17-1.40</td>
<td>4.22E⁻⁰⁸</td>
</tr>
<tr>
<td></td>
<td>5p12</td>
<td>rs5956944</td>
<td>G</td>
<td>1.31</td>
<td>1.39</td>
<td>1.34</td>
<td>1.19-1.50</td>
<td>4.75E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>1q13.1</td>
<td>rs8312277</td>
<td>T</td>
<td>1.24</td>
<td>1.26</td>
<td>1.25</td>
<td>1.15-1.36</td>
<td>5.34E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>8q21.3</td>
<td>rs12543318</td>
<td>C</td>
<td>1.31</td>
<td>1.16</td>
<td>1.25</td>
<td>1.15-1.36</td>
<td>2.11E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>8q22.2</td>
<td>rs1788160</td>
<td>G</td>
<td>1.35</td>
<td>1.20</td>
<td>1.29</td>
<td>1.17-1.41</td>
<td>9.52E⁻⁰⁸</td>
</tr>
<tr>
<td></td>
<td>8q24</td>
<td>rs55658222</td>
<td>A</td>
<td>2.08</td>
<td>1.91</td>
<td>2.00</td>
<td>1.78-2.26</td>
<td>5.63E⁻³⁰</td>
</tr>
<tr>
<td></td>
<td>9q21.31</td>
<td>rs6559624</td>
<td>T</td>
<td>1.45</td>
<td>1.19</td>
<td>1.34</td>
<td>1.20-1.50</td>
<td>1.72E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>10q26.3</td>
<td>rs7922405</td>
<td>A</td>
<td>1.24</td>
<td>1.25</td>
<td>1.25</td>
<td>1.14-1.46</td>
<td>4.61E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>15q24</td>
<td>rs28689146</td>
<td>A</td>
<td>1.24</td>
<td>1.35</td>
<td>1.28</td>
<td>1.17-1.40</td>
<td>5.40E⁻⁰⁸</td>
</tr>
<tr>
<td></td>
<td>17p13.1</td>
<td>rs11273201</td>
<td>AACCCAAAACCCAC</td>
<td>1.44</td>
<td>1.43</td>
<td>1.44</td>
<td>1.29-1.59</td>
<td>7.84E⁻¹⁲</td>
</tr>
<tr>
<td></td>
<td>19q13.11</td>
<td>rs7309428</td>
<td>A</td>
<td>1.65</td>
<td>1.31</td>
<td>1.50</td>
<td>1.30-1.73</td>
<td>2.92E⁻⁰⁸</td>
</tr>
<tr>
<td>European</td>
<td>1p36.13</td>
<td>rs9439714</td>
<td>C</td>
<td>1.62</td>
<td>1.38</td>
<td>1.52</td>
<td>1.30-1.89</td>
<td>1.91E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>6p21.33</td>
<td>rs79411602</td>
<td>C</td>
<td>1.67</td>
<td>1.31</td>
<td>1.52</td>
<td>1.29-1.78</td>
<td>2.92E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>8p24</td>
<td>rs72728734</td>
<td>G</td>
<td>2.22</td>
<td>1.84</td>
<td>2.04</td>
<td>1.70-2.44</td>
<td>7.33E⁻¹⁵</td>
</tr>
<tr>
<td></td>
<td>17p13.1</td>
<td>rs7406226</td>
<td>A</td>
<td>1.50</td>
<td>1.73</td>
<td>1.59</td>
<td>1.33-1.89</td>
<td>2.16E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>17q23.2</td>
<td>rs1588366</td>
<td>A</td>
<td>1.65</td>
<td>2.09</td>
<td>1.78</td>
<td>1.46-2.17</td>
<td>1.41E⁻⁰⁸</td>
</tr>
<tr>
<td>Asian</td>
<td>1q32</td>
<td>rs1044516</td>
<td>G</td>
<td>1.63</td>
<td>2.04</td>
<td>1.71</td>
<td>1.38-2.11</td>
<td>1.01E⁻⁰⁶</td>
</tr>
<tr>
<td>Central/ South American</td>
<td>1p22.1</td>
<td>rs4147882</td>
<td>T</td>
<td>1.55</td>
<td>1.41</td>
<td>1.48</td>
<td>1.27-1.73</td>
<td>3.99E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>1p32</td>
<td>rs11119345</td>
<td>G</td>
<td>1.85</td>
<td>1.77</td>
<td>1.81</td>
<td>1.55-2.13</td>
<td>3.40E⁻¹³</td>
</tr>
<tr>
<td></td>
<td>2q21</td>
<td>rs4850293</td>
<td>G</td>
<td>2.56</td>
<td>3.55</td>
<td>2.94</td>
<td>1.95-4.44</td>
<td>2.87E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>6p12.3</td>
<td>rs6918746</td>
<td>A</td>
<td>1.27</td>
<td>1.59</td>
<td>1.40</td>
<td>1.23-1.59</td>
<td>1.65E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>8q23.1</td>
<td>rs10780175</td>
<td>T</td>
<td>1.38</td>
<td>1.41</td>
<td>1.39</td>
<td>1.22-1.59</td>
<td>4.90E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>8q24</td>
<td>rs55658222</td>
<td>A</td>
<td>1.92</td>
<td>1.92</td>
<td>1.92</td>
<td>1.62-2.27</td>
<td>3.98E⁻¹⁴</td>
</tr>
<tr>
<td></td>
<td>16p13</td>
<td>rs111698437</td>
<td>T</td>
<td>2.33</td>
<td>3.39</td>
<td>2.74</td>
<td>1.86-4.03</td>
<td>3.67E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>17p13.1</td>
<td>rs7406226</td>
<td>A</td>
<td>1.60</td>
<td>1.28</td>
<td>1.45</td>
<td>1.28-1.65</td>
<td>1.46E⁻⁰⁸</td>
</tr>
<tr>
<td></td>
<td>17q22</td>
<td>rs227727</td>
<td>T</td>
<td>1.41</td>
<td>1.37</td>
<td>1.39</td>
<td>1.22-1.58</td>
<td>5.73E⁻⁰⁷</td>
</tr>
<tr>
<td></td>
<td>21q22</td>
<td>rs4920143</td>
<td>T</td>
<td>1.52</td>
<td>1.38</td>
<td>1.46</td>
<td>1.26-1.68</td>
<td>2.04E⁻⁰⁷</td>
</tr>
</tbody>
</table>
Our replication efforts focused on 10 SNPs in five loci: 2p24 (2 SNPs), 10q26.3 (2 SNPs) and 19q13 (3 SNPs) from the multi-ethnic analysis and 6p21 (1 SNP) and 17q23 (2 SNPs) from the European analysis. We genotyped 607 cases with CL/P and 1685 controls from three population-based case–control cohorts of European ancestry. The strongest evidence of association was with SNPs on 2p24 (rs7552, \( P = 2.68 \times 10^{-9} \); Table 2). Nominal \( P \)-values were observed for SNPs at the 17q23 locus (\( P = 0.02 \)). Although none of the three 19q13 SNPs were significant for CL/P, nominal \( P \)-values were observed for the CLP subgroup when stratified by cleft type.
several loci (3,9). The current study adds to this body of literature by identifying new risk loci (some with population specificity), by confirming the previously nominated loci in European ancestry, and also for the other Central/South American study subjects. Therefore, it is particularly notable that the Central/South American group showed evidence of association with both OFC GWAS peaks previously identified with either European or Asian ancestry.

Overall, the results of our multi-ethnic study corroborate findings from previous GWAS. Fifteen loci had reached genome-wide significance in the previous six GWASs or meta-analysis (3–6,8,9). Supplementary Material, Table S8 lists the P-values from the present study for the previously published SNPs. As discussed above, five of these loci reached genome-wide significance in this study. Five additional loci showed suggestive evidence of association (8q21 near MMP16, 10q25 near VAX1, 13q31 near SPRY2, 17q22 near NOG and 20q12 near MAFB). Weaker evidence for association was observed at three loci: 2p21 near THADA, 3p11 near EPHA3 and 16p13 near CRBBP. In aggregate, our results support a majority of the previously reported risk loci and implicate additional novel loci, indicating that the etiology of CL/P is highly heterogeneous.

**Discussion**

GWASs have allowed for the identification of risk loci for complex diseases and traits. For CL/P, six GWASs and a meta-analysis have identified 15 loci with genome-wide significance, confirming the multifactorial nature of this disease (3–6,8,14). In addition, these efforts have revealed population specificity of several loci (3,9). The current study adds to this body of literature by identifying new risk loci (some with population specificity), by confirming the previously nominated loci in European and Asian populations, and by expanding the GWAS approach to African and Central/South American populations.

**Insights into new risk loci**

19q13. SNPs at the new 19q13 risk locus spanned RHPN2. This gene encodes the ubiquitously expressed RhoA binding protein, Rhophilin 2 (19). Although there are few studies devoted to elucidating the function of RHPN2, biochemical assays suggest that its normal function is to decrease stress fibers, a component of the actin cytoskeleton (19). RhoA belongs to a family of precisely regulated molecular switches that modulate a variety of cellular processes such as cytoskeletal reorganization, cell migration and cell–cell interactions. Several lines of evidence point to the critical role of Rho signaling and cell migration in craniofacial development. First, Irf6-deficient keratinocytes exhibit delayed migration, prominent stress fibers and increased RhoA activity (20). In contrast, Grhl3-deficient keratinocytes have decreased stress fibers and RhoA activity (21). Both Irf6 and GRHL3 are mutated in Van der Woude syndrome, the most common OFC syndrome (22,23). Activation of RhoA is required for transforming growth factor beta 3 signaling during palate formation (24). We previously investigated the 1p22 GWAS locus with expression studies and sequencing that implicated ARHGAP29 (a RhoA GTPase activating protein) as the etiologic OFC gene (25). Although additional studies into the role of RHPN2 in craniofacial development are needed, these association results are additional evidence pointing to abnormal regulation of Rho activity as a pathophysiological mechanism for OFCs.

17q23. SNPs in the 415 kb associated region on 17q23 span several genes including TANC2, CYB561 and a hypothetical micro-RNA. A definitive role for TANC2 in craniofacial development is not immediately apparent; however, mice homozygous for a gene trap allele die prior to embryonic day 12, suggesting that this gene is required for development (26). In addition, TANC2 is an oncogene that regulates epithelial cell growth and survival (27). If future studies confirmed a role in craniofacial development, TANC2 would join of the growing list of genes in which allelic variants increase the risk for certain cancers and also confer risk for OFCs including IRF6 (28), CDH1 (29) and FOXE1 (30). Another compelling candidate in the 17q23 region is DCAF7, located ~500 kb from the peak signal. DCAF7, also known as WDR68, encodes a conserved Trp-Asp repeat protein that is thought to serve as a scaffold to facilitate protein–protein interactions. In zebrafish, wdr68 is required for craniofacial development (31,32). The dys mutant, which harbors an insertion in this gene, has reduced Meckel’s and palatoquadrate cartilages, which serve as the embryonic lower and upper jaws, respectively (31). This phenotype resembled that of endothelin-1 (edn1) mutants leading to the conclusion that wdr68 is required for all edn1-dependent cartilages of the first arch. Although a Dcaf7+/− mouse has not been generated, Edn1−/− mice have many craniofacial anomalies including auricular and mandibular hypoplasia and CP (33). Moreover, micrognathia, bifid uvula
and other oral anomalies have been described in humans with homozygous EDN1 mutations (34). Together, these data suggest that DCAF7/WDR68 and endothelin signaling have conserved roles in craniofacial development in vertebrates. Although this gene is located far from the peak signal, common etiologic variants could be located in regulatory elements of DCAF7, as has been shown for other OFC risk loci (13,25,30).

2p24. The associated SNPs at the 2p24 locus span the 3’ untranslated region of FAM49A and its downstream region. Little is known about FAM49A, which is expressed in the brain in mouse and zebrafish (35,36) and encodes a conserved protein of unknown function. Another compelling gene in the region is MYCN, the gene mutated in Feingold syndrome. This syndrome is characterized by digital anomalies, microcephaly and facial dysmorphism that can include OFCs.

Among the second tier hits, two loci (5p12 and 15q24) included compelling putative OFC risk genes. The 5p12 locus contains FGFP and FGF10, both of which have been implicated in palatogenesis in humans and mice (37,38). Our peak signal at the 15q24 locus is within the 1.1-Mb critical region for 15q24 palatogenesis. This syndrome is characterized by digital anomalies, microcephaly and facial dysmorphism and CF among the described features (39).

Conclusion

The goals of this study were to identify novel risk genes for CL/P and to extend the GWAS approach to understudied populations, such as admixed populations from Central and South America. The large sample sizes allowed us to accomplish both of these goals. In addition to confirming previous findings for nine risk loci, we identified three new risk loci with genome-wide significance and many other loci with compelling statistical support. We also confirmed the population specificity of some risk loci (e.g. 8q24 in Europeans) and showed that risk loci from both European and Asian groups were associated in the Central/South American group although no genome-wide significant novel risk loci specific to this admixed group were observed. Taken together, these results advance our understanding of the genetic architecture of OFCs, highlight the importance of a multi-ethnic approach to this common birth defect and may ultimately prove useful for predicting recurrence and prognosis.

Materials and Methods

Cohort description

The cohort for this study comes from a worldwide sample totaling 11,727 participants recruited from 18 sites across 13 countries from North America, Central or South America, Asia, Europe and Africa. Most of the sites were a part of ongoing genetic and phenotyping studies led by the University of Pittsburgh Center for Craniofacial and Dental Genetics; some sites were also a part of ongoing genetic studies led by the University of Iowa. In general, recruitment and study assessments were done at regional centers for OFC treatment. All sites had Institutional Review Board approval both locally and at the University of Pittsburgh or University of Iowa as appropriate for participation after informed consent and for large-scale genomic studies including data sharing.

The overall study cohort includes OFC-affected probands, their unaffected family members and controls with no known history of OFC or of other craniofacial anomalies. The specific affection status included for the current study was CL with or without CP (CL/P). We partitioned the total sample into two mutually exclusive analysis sets for the current study: (1) a subset of 1319 case–parent trios (i.e. 3957 individuals; note, from each multiplex family only one trio was chosen), and (2) a subset of 823 unrelated CL/P cases and 1700 unrelated controls (Supplementary Material, Table S1). Note that these unrelated cases did not have DNA available for one or both parents, and further that there is no overlap between the trios and the case-control group. See the flow chart depicted in Supplementary Material, Figure S2 showing the process for assigning participants into analysis subset. A total of 317 individuals (35 cases, 282 individuals in case–parent trios) were previously genotyped and included in the CL/P GWAS by Beaty et al. (3). All Guatemalan individuals were previously genotyped and included in the GWAS by Wolf et al. (14).

Genotyping, quality control and imputation

Samples were genotyped for ~580K SNPs using an Illumina platform by the Center for Inherited Disease Research (CIDR) at Johns Hopkins University, and data quality assurance and data cleaning were performed in collaboration with the CIDR Genetics Coordinating Center (GCC) at the University of Washington using analysis pipelines developed by the GCC (40). A total of 539,473 genotyped SNPs (comprising 96.74% of those attempted) passed quality filters recommended by the GCC (Supplementary Material, Table S2). Of these, 295,833 SNPs had a minor allele frequency (MAF) of ≥1%. Imputation of 34,985,077 unobserved genetic polymorphisms was conducted using the software IMPUTE2 (41) and phase 3 of the 1000 Genomes Project (comprising 2504 individuals from 26 populations worldwide) as the reference panel. SNPs passing quality filters were ‘pre-phased’ to create haplotypes using SHAPEIT2 software (42) prior to imputation. Imputation accuracy was assessed by masked variant analysis, demonstrating high-quality imputation, with mean concordance of 0.995 for SNPs with MAF < 0.05 and 0.960 for SNPs with MAF ≥ 0.05. The ‘most-likely’ genotypes (i.e. genotypes with the highest probability (Q)) were selected for statistical analysis if and only if the highest probability > 0.5. Imputed SNPs with deviations from Hardy-Weinberg equilibrium in European controls were filtered out of downstream analyses. Further details regarding the quality assurance, ancestry analyses, imputation, and data cleaning efforts are available in the Quality Control Report issued by the University of Washington GCC for this project and are available online (http://www.ccdg.pitt.edu/docs/Marazita_ofc_QC_report_feb2015.pdf, last accessed April 25, 2016).

PCA

PCA analysis was used, incorporating identity-by-descent estimates calculated for all pairs of participants (43) to extract the maximal subset of unrelated participants who were then included in PCA (using a pruned set of 67K SNPs chosen to be in low LD). The resulting eigenvectors were projected onto the remaining set of relatives. Supplementary Material, Figure S1 shows plots of the PCs, which strongly tracked global recruitment site and self-reported race/ethnicity. Further, although PCs of ancestry represent orthogonal information across the total multi-ethnic sample, individual PCs may be highly collinear in specific strata. Therefore, in addition to the cosmopolitan PCs generated across all participants, PCs of ancestry were also generated within each continental group (European, Central/South...
American and Asian, as defined by cosmopolitan PCs) for use in stratified analyses.

Statistical analyses
The TDT (44) was used to perform a GWAS in the trio set. In brief, this test detects evidence of over-transmission of a particular allele from parents to affected offspring and is robust to any population stratification. The GWAS in the unrelated case-control analysis set was performed using logistic regression while adjusting for the first 18 PCA in order to protect against genomic inflation due to population structure. Both the TDT test and the logistic regression analyses were implemented in PLINK (45). The weighted OR method of meta-analysis (46) was used to combine TDT and case-control results. We performed a GWAS in the total sample, and also stratified by continental ancestry group (European, Central/South American and Asian, but not African due to insufficient sample size). For stratified analyses, we adjusted for the first five PCs in European and Central/South American samples and first three PCs in the Asian sample.

We used the conventional threshold for declaring genomewide significance, \( P \)-value of \( 5 \times 10^{-8} \), corresponding to the Bonferroni correction for 1 million tests. Furthermore, as an exploratory effort we scrutinized ‘suggestive’ associations with \( P \)-values \( <5 \times 10^{-7} \) for nearby (i.e. physically proximal) genes with known biological functions that may be involved in craniofacial development. Calculations of the genomic inflation factor, \( \lambda \), showed no significant inflation. Results of these analyses will be available for browsing as a part of the Facebase Consortium’s Human Genomics Analysis Interface Tool (www.facebase.sdmgenetics.pitt.edu, last accessed April 25, 2016).

Replication cohort and genotyping
Samples utilized for replication genotyping came from five OFC case-control population-based studies (two American and three European): the Danish National Birth Cohort (47), the Iowa Case-Control Study (Iowa) (48), the Norway Facial Clefts Study (49), the Norwegian Mother and Child Cohort (MoBa) (50), and the Utah Child and Family Health Study (Utah) (51) (Supplementary Material, Table S3). Cases and controls were identified primarily through national or state-based birth registries. Ten SNPs were selected for replication. If the best genotyped or imputed SNP was not available, a surrogate SNP was chosen. Genotyping was performed using TaqMan SNP Genotyping Assays (Life Technologies, Carlsbad, CA) on the Fluidigm microfluidic EP1 SNP Genotyping System and GT192.24 Dynamic Array Integrated Fluidic Circuits (Fluidigm, San Francisco, CA). We considered replication \( P \)-values \( <0.005 \) to be statistically significant (0.05/10 SNPs).

Supplementary Material
Supplementary Material is available at HMG online.

Acknowledgements
Authors thank the participating families who made the studies summarized here possible, and who provided on-going inspiration to the investigators. The tireless efforts by the dedicated field staff and collaborators was similarly essential to the success of this study. Many thanks to Kathy Bardi, Rebecca DeSensi, Margaret Cooper, Joel Anderton, Wendy Carricato, Kathleen Deeley, Keri Simonette, Matthew Ford, and Dr. Joseph Losee (University of Pittsburgh); Jennifer Standley, Nichole Nidey, Camden Bay and Erin Brothers-Smith (University of Iowa); Marcia Adams, Marie Hurley and Mary Jewell (Center for Inherited Disease Research, Johns Hopkins University); András Kiss and Eszter Nyulátsi (Hungary); Rosemary Vance and the Children of the Americas (www.childrenoftheamericas.org, Guatemala); Elena Serna, Rosa Martinez, and Syed S. Hashmi (University of Texas, Houston); Dr. Ajit Ray (India); Dr. You-e Liu (China); Dr. Rick Martin (Missouri); Sondra Valdez (University of Colorado); the Smile Train and Transforming Faces Worldwide (Africa).

Conflict of Interest statement. None declared.

Funding
National Institutes of Health (R01-HG007485 to MLM, R01-DE016148 to M.L.M., U01-DE024425 to M.L.M., R37-DE008559 to J.C.M. and MLM, R01-DE009886 to M.L.M., R21-DE016930 to M.L.M., R01-DE014667 to A.C.L. and L.M.M., R21-DE016930 to M.L.M., R01-DE012472 to M.L.M., R01-DE011931 to J.T.H., R01-DE011948 to K.C., R01-DO000295 to G.L.W., U54-DO007587 to C.J.B., R25-MD007607 to C.J.B., K99-DE024571 to C.J.B., K99/R00-DE022378 to A.B., K99-DE025060 to E.J.L., and R01-DE020895 to G.L.W.). Genotyping and data cleaning were provided via an National Institutes of Health contract HHSN268201200008I to the Johns Hopkins Center for Inherited Disease Research. Additional support provided by: the Robert Wood Johnson Foundation, AMFDP Grant (72429 to A.B.); an intramural grant from the Research Institute of the Children’s Hospital of Colorado (FWD); operating costs support in the Philippines was provided by the Institute of Human Genetics, National Institutes of Health, University of the Philippines, Manila (C.P.); grants through Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, Brazil (I.M.O.); grant numbers (E-26/102.797/2012, E-26/110.140/2013); grants through CNPq, Brazil (I.M.O.); grant numbers: (481069/2012-7, 506396/2013-0, 400427/2013-3). Some replication samples were from the Utah Child and Family Health study (R01-DE016877 [R.G.M.]. Some replication samples were from the Danish National Birth Cohort (Pls Mads Melbye and Jørn Olsen) funded by a major grant from the Danish National Research Foundation and initiated by the Danish Epidemiology Science Centre. Additional support for the Danish National Birth Cohort was obtained from the Pharmacy Foundation, the Egmont Foundation, the March of Dimes Birth Defects Foundation, the Augustinus Foundation, and the Health Foundation. Some replication samples were from the Norwegian Cleft Study and from the Norwegian MoBa Study, both supported by the National Institute of Environmental Health Sciences/National Institutes of Health Intramural Research Program and the Norwegian Research Council.

References


