

Pleiotropy method identifies genetic overlap between orofacial clefts at multiple loci from GWAS of multi-ethnic trios

Debashree Ray^{1,2,✉}, Sowmya Venkataraghavan¹, Wanying Zhang¹, Elizabeth J. Leslie³, Jacqueline B. Hetmanski¹, Mary L Marazita^{4,5,6}, Ingo Ruczinski², Margaret A. Taub² and Terri H. Beaty^{1,2,✉}

¹*Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA.*

²*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA.*

³*Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA.*

⁴*Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, PA, USA.*

⁵*Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, PA, USA.*

⁶*Center for Craniofacial and Dental Genetics, University of Pittsburgh, PA, USA.*

✉Email: dray@jhu.edu; tbeaty1@jhu.edu

Based on epidemiologic and embryologic patterns, nonsyndromic orofacial clefts are commonly categorized into cleft lip with or without cleft palate (CL/P) and cleft palate alone (CP). While nearly forty risk genes have been identified for CL/P, few risk genes are known for CP. We used a new statistical method, PLACO, to identify genetic variants influencing risk of both CL/P and CP. In a combined multi-ethnic genome-wide study of 2,771 CL/P and 611 CP case-parent trios, we discovered 6 new loci of genetic overlap between CL/P and CP; 3 new loci between pairwise OFC subtypes; and 4 loci not previously implicated in OFCs. We replicated the shared genetic etiology of subtypes underlying CL/P, and further discovered loci of genetic overlap exhibiting etiologic differences. In summary, we found evidence for

new genetic regions and confirmed some recognized OFC genes either exerting shared risk or with opposite effects on risk to OFC subtypes.

1 INTRODUCTION

2 Orofacial clefts (OFCs) are the most common craniofacial birth defects that severely affect financial
3 and psychological well-being and the overall quality of life of the affected child and their family¹.
4 These malformations most commonly occur as isolated defects (i.e., nonsyndromic clefts) and
5 affect, on average, nearly 1 out of 1,000 live births worldwide². People born with OFCs require
6 multi-disciplinary medical treatments; have increased risk of psychological problems³; have greater
7 risk of various types of cancer (e.g., breast, brain and colon)⁴; and have increased mortality
8 throughout the life course⁵. Overall, OFCs pose a major public health burden, with underlying
9 biological mechanisms largely unknown.

10 Nonsyndromic OFCs typically manifest as a gap in the upper lip ('cleft lip' or CL) or the roof
11 of the mouth ('cleft palate' or CP) or both ('cleft lip and palate' or CLP). Based on epidemiologic
12 evidence, prevalence rates and the embryologic period when they develop, the subtypes CL and
13 CLP are typically grouped together as the subgroup CL/P (cleft lip with or without palate)⁶⁻⁸, while
14 CP alone forms the other subgroup. CL/P and CP have been historically analyzed separately⁹⁻¹².
15 While genetic studies have identified nearly 40 genetic regions (or loci) as significantly associated
16 with risk to CL/P, fewer, around 10 loci, have been identified for CP^{2,13-15}. The findings for CP have
17 mostly been identified in the Han Chinese population¹⁵. Together, these genetic regions explain no
18 more than a quarter of the estimated total heritability of risk to OFCs¹⁶.

19 Although the OFC subgroups CL/P and CP have been considered distinct, shared genetic
20 risk variants have been suggested^{9,11}. There are multiplex cleft families with both CL/P and CP
21 present in affected relatives^{17,18}. In recent years, there have been attempts to discover overlapping
22 genetic etiology of OFC subtypes. In this context, it is important to distinguish between genetic

23 overlap and genetic heterogeneity. While genetic heterogeneity may refer to shared genetic effects
 24 as well as subtype-specific effects (which may mean a non-null effect on one subtype and no effect
 25 on the other), genetic overlap refers to non-null genetic effects on both subtypes that may or may
 26 not be equal in magnitude and/or direction. The usual approach for identifying genetic overlap in
 27 OFCs is to compare the significant findings from one subtype with those from the other^{11,13,19,20}.
 28 However, the discovery of the associated variants in the first place may be under-powered in
 29 genome-wide association studies (GWAS) of each subtype separately. For instance, success of
 30 discovery genetics has been elusive for CP, which could partially be due to smaller sample sizes of
 31 CP¹⁹ reflecting its lower birth prevalence. Another approach involves testing how well polygenic
 32 risk scores for one subtype can explain variation for another^{13,21}, which describes overall genetic
 33 sharing, does not implicate specific regions of overlap (novel or otherwise), and may indicate lack
 34 of overlap when one subtype has a much smaller sample size^{13,14}. One strategy is the ‘pooled
 35 method’ GWAS analysis¹², where all the OFC subtypes are pooled together in a combined analysis
 36 of all OFCs. *FOXE1* has been successfully implicated as a shared risk gene using this approach¹².
 37 While association signals from the pooled method may be driven by shared risk variants between
 38 subtypes, the pooled method does not necessarily capture only shared signals²², especially if
 39 sample sizes are widely different between the subtypes (e.g., the CL/P group is almost always
 40 much larger than the CP group) or if strong genetic effects exist in one group but not the other.
 41 Furthermore, if a locus is hypothesized to have opposite genetic effects on the two subtypes (e.g.,
 42 *NOG*^{14,19}), the pooling technique will dilute any signal and consequently will be under-powered
 43 to detect genetic overlap. Use of multi-trait methods in OFC genetic studies, such as the ones
 44 commonly used in population-based GWAS of complex traits^{23–26}, is hindered by the disjoint
 45 nature of the subtypes (i.e., absence of subjects with both traits), the qualitative (binary) nature
 46 of the traits, and/or the case-parent trio design typically used to study multi-ethnic samples with
 47 OFC.

In this article, we use a new statistical method for pleiotropic analysis under composite null hypothesis, PLACO, to discover genetic variants influencing risk of the two major nonsyndromic OFC subgroups (CL/P & CP). Although PLACO was originally developed to discover pleiotropic variants between two traits from population-based studies²⁷, we found it can also help identify genetic variants simultaneously influencing risk in two disease subgroups from family-based studies (see **Methods**). PLACO is particularly useful and powerful in identifying variants that increase risk of one subgroup while decreasing risk for the other, and seems to be robust to modest difference in sample sizes and in effect sizes between subgroups. We performed a meta-analysis GWAS using PLACO on 2,771 CL/P and 611 CP multi-ethnic case-parent trios from the Pittsburgh Orofacial Cleft (POFC) and the Genes and Environment Association (GENEVA) studies. To dissect the genetic architecture at regions of genetic overlap between CL/P & CP, we also explored genetic overlap stratified by racial/ethnic group, investigated if the overlapping genetic etiology is modified by sex, and explored if our findings are driven by specific pairs of OFC subtypes (CL & CP or CLP & CP).

RESULTS

Identification of 9 loci with genetic overlap between CL/P & CP, including 2 novel loci for OFCs. At the genome-wide significance level 5×10^{-8} , PLACO identified 1 locus in a well-recognized risk gene that also happens to be a candidate shared gene²⁸, 1q32.2 (*IRF6*, $p = 4.3 \times 10^{-12}$), while 2 loci in 1p36.13 (*PAX7*, $p = 6.9 \times 10^{-8}$) and 17q22 (*NOG*, $p = 6.0 \times 10^{-8}$) are suggestive, barely missing this significance threshold (**Figure 1**). Additionally, 6 loci showed evidence for genetic overlap between CL/P & CP at a suggestive threshold of 10^{-6} : 3q29 (*DLG1*, $p = 5.3 \times 10^{-7}$), 4p13 (*LIMCH1*, $p = 5.0 \times 10^{-7}$), 4q21.1 (*SHROOM3*, $p = 8.1 \times 10^{-7}$), 9q22.33 (*FOXEL1*, $p = 1.7 \times 10^{-7}$), 19p13.12 (*RAB8A*, $p = 6.8 \times 10^{-7}$) and 20q12 (*MAFB*, $p = 9.9 \times 10^{-7}$). The 2 loci in *LIMCH1* and *RAB8A* are novel for OFCs. All the other genes have been implicated in GWAS of CL/P previously^{2,14}, and insights into the molecular pathogenesis of OFCs via many

of these genes is summarized elsewhere²⁹. QQ plots from all our analyses show deviation from the null only in the tail end of the distribution of p-values (**Figures S1, S2**), indicating genetic signals rather than any systemic bias.

Six out of 9 loci yielding novel evidence for genetic overlap between CL/P & CP. Of the 9 loci, genetic overlap at SNPs in/near genes *IRF6*²⁸, *FOXE1*¹² and *NOG*¹⁹ have been previously suggested in GWAS of clefts. We found novel, strong statistical evidence of genetic overlap at the 6 loci in/near *PAX7*, *DLG1*, *LIMCH1*, *SHROOM3*, *RAB8A* and *MAFB*. In particular, PLACO provided stronger evidence for a pleiotropic association compared to the marginal association of each subtype for these markers in *DLG1*, *LIMCH1* and *RAB8A* loci (**Table 1**).

Genetic sharing at these loci are not uniform in their effect on risk. We found the chosen effect alleles at the lead SNPs in/near *FOXE1*, *RAB8A* and *MAFB* loci appear to increase risk for both CL/P and CP, while the effect alleles at the remaining loci affect these OFC subgroups in opposing directions as reflected by the estimated relative risks (RR) (**Table 1**). In other words, the effect alleles at the lead SNPs at *PAX7*, *IRF6*, *DLG1*, *LIMCH1*, *SHROOM3* and *NOG* seem to predispose to one OFC subgroup while protecting from the other. The estimated RRs of the top several SNPs at each of these loci further support this finding (**Figure 2**). In particular for markers in the *LIMCH1*, *IRF6*, *DLG1* and *NOG* loci, the estimated RRs of the lead SNP for subtypes CL and CLP (and hence CL/P) and their corresponding 95% confidence intervals (CIs) were all completely on the same side of the null value 1, while the estimated RR and its 95% CI for CP was completely on the other side (**Figures 3, S3, S4, S5; panel b**). The opposite effects of these effect alleles likely explain why these loci were not conclusively identified as influencing risk to both CL/P and CP in the ‘pooled method’ GWAS analysis of all OFC subtypes from POFC and GENEVA subjects before¹². Additionally, the *IRF6* region appears to harbor at least 2 distinct loci: one with shared genetic effects, and another with opposite effects (**Figures 2, S6**). Evidence for

97 both shared and opposite effects at *IRF6* has been reported previously^{14,30}.

98 **Genetic overlap between subgroups CL/P & CP is consistent with overlap identified between**
 99 **pairwise OFC subtypes CL & CP and CLP & CP.** To gain a better understanding of which
 100 subtypes are driving this evidence of genetic overlap between CL/P & CP, we applied PLACO on
 101 the pairwise component OFC subtypes (**Figure 4**). The *PAX7*, *SHROOM3*, *FOXEL* and *MAFB* loci
 102 seem to be driven by the common genetic basis of subtypes CLP & CP at these loci (**Figures S7,**
 103 **S8, S9, S10**; panel **g**). The rest of the loci (*LIMCH1*, *IRF6*, *DLG1*, *NOG* and *RAB8A*) appear to
 104 be driven by genetic overlap between CL & CP as well as CLP & CP (**Figures 3, S3, S4, S5, S11**;
 105 panels **f, g**).

106 **Identification of additional novel regions of genetic overlap between component OFC subtypes.**
 107 PLACO revealed 2 loci in 18q12.1 (*MIR302F*, $p = 6.2 \times 10^{-7}$) and 10q24.33 (*SH3PXD2A*,
 108 $p = 9.2 \times 10^{-7}$) associated with CL & CP, and CLP & CP subtypes respectively at a suggestive
 109 threshold of 10^{-6} (**Table 1**). Effect alleles at the lead SNPs of both these loci increase risk for
 110 one subtype while decreasing risk for the other (**Figures S12, S13**; panel **b**). The RR estimates
 111 and their 95% CIs for the top several SNPs at these loci confirmed the opposite effect of these
 112 loci on risks of OFC subtypes (**Figure 2**). While *SH3PXD2A* has been recently implicated in the
 113 formation of CL/P in an European GWAS³¹, the *MIR302F* locus is a novel genetic risk factor for
 114 OFCs.

115 **Locus-specific effects at regions of genetic overlap between CL/P & CP vary by racial/ethnic**
 116 **group.** The regional association plots seem to indicate that signals of genetic overlap at the *PAX7*,
 117 *FOXEL* and *NOG* loci are driven by the European subjects; *IRF6*, *SHROOM3* and *MAFB* loci by
 118 the Asian subjects; while the *DLG1* and *RAB8A* loci seem to draw upon evidence from both groups
 119 (**Figures S3-S11**; panels **c-e**). Similarly, evidence for the *LIMCH1* locus seems to be driven by

both Asian and Latin American subjects (**Figure 3**; panels **c, e**). However, we must note that signals in these stratified analyses are confounded by differences in overall sample sizes between racial/ethnic groups (**Table S1**), sample size distribution between CL/P and CP subgroups, as well as minor allele frequency (MAF) differences across racial/ethnic groups. Consequently, the regional association plots do not fully indicate the differential information content of SNPs across these racial/ethnic groups. We, therefore, additionally provide a forest plot of RR estimates from the genotypic transmission disequilibrium test (gTDT) analyses stratified by cleft subtype and by racial/ethnic group for the lead SNP from each common locus (**Figures 3, S3-S11**; panel **b**). Notably for the Latin American subjects, the large uncertainty in the estimates for CP and the hugely skewed ratio of sample sizes between CL/P and CP are quite evident, leading to lack of power to drive signals of genetic overlap in this stratified analysis. The Asian and the European subgroups are more comparable in size, and findings from the regional association plots of these two racial/ethnic groups seem to be reflected in the forest plots as well.

These regions of genetic overlap do not appear to be modified by sex. There is increasing evidence for sex-specific differences in human health and disorders. While CL/P is 2 times more common in males, CP is more common in females^{2,29}. Recent studies have indicated sex-specific differences in pleiotropic effects on complex traits³². We used PLACO to test for non-null SNP×Sex interaction effects in both CL/P & CP, and failed to find any statistical evidence of pleiotropic effect modification by sex at the 9 loci of genetic overlap (**Figure S14**). We also did not find effect modification by sex at the 2 loci of genetic overlap, 18q12.1 and 10q24.33, identified between specific OFC subtypes. Tests of statistical interaction require larger sample sizes than tests of main effects; perhaps our sample size is not large enough to identify any effect modification by sex.

Proof of principle for sensitivity of PLACO in discovering shared etiology between OFC subtypes. Analysis of subtypes CL & CLP using PLACO identified nearly all the recognized

144 risk genes for CL/P subgroup alone as detected by the conventional gTDT analysis (**Figure 5**).
 145 In other words, the regions of genetic overlap identified by PLACO matched the shared signals
 146 captured by the pooled method analysis of CL and CLP subtypes. The RR estimates in **Table**
 147 **S2** indicate the effect alleles at lead SNPs in all regions of genetic overlap affect risk of both CL
 148 and CLP in the same direction, which is consistent with the vast literature of epidemiologic and
 149 genome-wide studies of OFCs^{6-8,12,14,20,30}.

150 **Beyond shared etiology, few loci of genetic overlap exhibit etiologic differences between**
 151 **CL and CLP.** When investigating the RR estimates from the top several SNPs at each of the
 152 above-mentioned loci shared by CL and CLP, we found loci suggesting different pathogenesis
 153 of these subtypes at 1p22.1 (*ABCA4*, *ARHGAP29*) and 1q32.2 (*IRF6*) as evidenced by a large
 154 number of SNPs with opposite genetic effects (**Figure S15**). Some previous studies^{14,33} have noted
 155 differences in genetic etiologies of CL and CLP, particularly at the *IRF6* locus. Additionally,
 156 the 1p36.13 (*PAX7*), 3q12.1 (*COL8A1*), 8q21.3 (*DCAF4L2*) and 8q24 (gene desert) regions with
 157 shared effects between CL and CLP appear to have at least 2 distinct loci of genetic overlap
 158 indicated by more than one peaks (**Figure S16**). Presence of possibly independent loci in the 8q24
 159 region has been reported previously¹⁴. Furthermore, PLACO revealed a novel OFC locus at 1p21.3
 160 (*MIR137HG*, $p = 2.2 \times 10^{-8}$) with opposite genetic effects for CL & CLP at the genome-wide
 161 threshold (**Table 1** and **Figure S17 a**). The estimated RRs of the effect allele at the lead SNP
 162 of this locus indicate its protective effect on CL, and its deleterious effect on subtypes CLP and
 163 CP across racial/ethnic groups (**Figure S17 b**). This signal is, however, lost from CL/P (**Figure**
 164 **S17 c**) due to pooling together of opposite effects of variants on CL and CLP (**Figure S17 d, e**).
 165 Consequently, this locus near *MIR137HG* fails to show any evidence of genetic overlap between
 166 CL/P & CP (**Figure S17 f**) even though there is moderately strong statistical evidence of genetic
 167 overlap between all pairs of OFC subtypes (**Figure S17 a, g, h**).

In-silico validation of robustness of PLACO to differences in sample sizes and modest subgroup-

specific effects. To appreciate the advantages of PLACO and to interpret the following empirical results, it is important to briefly describe the intuition and statistical model behind PLACO. For two disease subgroups with genetic effects $\beta_1 = \log(\text{RR}_1)$ and $\beta_2 = \log(\text{RR}_2)$ for a variant, three possible situations can arise: global null, where the variant has no genetic effect on either subgroup ($\beta_1 = 0, \beta_2 = 0$); sub-null, where the variant influences risk to one subgroup but not the other (either $\beta_1 = 0, \beta_2 \neq 0$ or $\beta_1 \neq 0, \beta_2 = 0$); and finally non-null, where the variant influences risk to both subgroups ($\beta_1 \neq 0, \beta_2 \neq 0$). Only the non-null situation here describes genetic overlap between the two subgroups. PLACO tests a composite null hypothesis comprising both the global null and the sub-null situations, and thus rejection of this composite null provides statistical evidence of genetic overlap at a given variant (see **Methods**). On the other hand, the pooled method (previously used to identify risk variants common to both CL/P and CP)¹² tests the global null hypothesis, and it may be rejected because either the sub-null or the non-null situations exist.

We compared PLACO with the pooled method across multiple simulation scenarios. When almost all of the simulated null variants had no effect on either OFC subgroup (Scenario I – majority of variants under the global null situation), both PLACO and the pooled method showed well-controlled type I error rates (**Figure S19 a**). As the sample sizes became skewed between the OFC subgroups, the pooled method showed inflated type I error while PLACO maintained appropriate type I error rate even at stringent error levels (**Figure S19 b-c**). When a large proportion of simulated null variants had a genetic effect on one OFC subgroup only (Scenario II – majority of variants under the sub-null situation), the pooled method had hugely inflated type I error while PLACO showed proper type I error control at stringent levels regardless of skewed sample sizes between the two OFC subgroups (**Figure S20 a-c**). This observation holds true irrespective of how widely different the MAFs are between the two simulated ethnic groups (**Figure S20 d-f**). This

shows how the pooled method may show spurious signals (or increased false discoveries) if genetic effect exists in one OFC subgroup but not the other, and if there is a large sample size difference between subgroups. On the other hand, our empirical results suggest robustness of PLACO's type I error to sample size differences between OFC subgroups; moderately strong subgroup-specific effects; and small to large MAF differences between ethnic groups.

Empirical evidence of sensitivity of PLACO in discovering common genetic basis of OFC subgroups. We benchmarked the power of PLACO against the pooled method (even though pooled method shows increased false discoveries under the sub-null situations) along with the naive approach of declaring genetic overlap when a variant reaches genome-wide significance for the larger OFC subgroup (in our case, CL/P) and reaches a more liberal significance threshold for the other. We used two such naive approaches: one based on the criterion $p_{CL/P} < 5 \times 10^{-8}$, $p_{CP} < 10^{-5}$ and the other $p_{CL/P} < 5 \times 10^{-8}$, $p_{CP} < 10^{-3}$ ('Naive-1' and 'Naive-2' respectively in our figures). Regardless of the magnitude and directions of genetic effect on these OFC subgroups, and the sample size differences, PLACO showed dramatically improved statistical power to detect common genetic basis compared to the naive approaches (**Figure S21**). For instance, the 'Naive-2' method with a liberal threshold criterion has a 24% power, compared to 61% for PLACO, to detect simultaneous association using 1,800 CL/P and 600 CP trios when an MAF 10% variant influences risk to both CL/P and CP with $RR = 1.5$. This probably explains why a genome-wide analysis of CL/P first, followed by an analysis of CP on the most significant findings, have not quite proven successful in providing evidence of overlapping association signals between these two OFC subgroups. Although PLACO was slightly less powerful than the pooled method in identifying shared risk variants, it did achieve greater power gain when detecting variants that increased risk for one OFC subgroup while decreasing risk for the other (**Figure S21**).

216 DISCUSSION

217 In this analysis of multi-ethnic case-parent trios from the POFC and the GENEVA studies, we
 218 identified genetic overlap between nonsyndromic CL/P & CP at 1 locus in 1q32.2 reaching genome-wide
 219 significance (5×10^{-8}), 2 loci in 1p36.13 and 17q22 barely missing this conventional threshold,
 220 and 6 loci in 3q29, 4p13, 4q21.1, 9q22.33, 19p13.12 and 20q12 yielding suggestive significance
 221 at a threshold of 10^{-6} . The apparent risk SNPs at 4p13 and 19p13.12 are located in the *LIMCH1*
 222 and *RAB8A* genes respectively, which have not been implicated in OFC genetics before. We found
 223 evidence of shared etiology at 3 of these 9 loci, including the well-recognized *FOXE1* gene which
 224 influences risk to both CL/P and CP in the same direction. The effect alleles at the other loci
 225 found in this analysis appear to increase risk for one OFC subgroup but decrease risk for the other.
 226 We also identified genetic overlap between CL & CP, CLP & CP, and CL & CLP at 3 loci in
 227 18q12.1, 10q24.3 and 1p21.3 respectively, of which the loci 18q12.1 near *MIR302F* and 1p21.3
 228 near *MIR137HG* have not been previously shown to be associated with risk to OFCs. We replicated
 229 shared etiology of CL & CLP in/near several recognized genes, and further found opposite effects
 230 of top several SNPs at a few loci hinting at potentially different pathogenesis of these 2 OFC
 231 subtypes. None of the loci identified in this study appear to exhibit sex-dependent genetic overlap.

232 While *LIMCH1*, *RAB8A*, *MIR302F* and *MIR137HG* are novel risk genes for OFCs, all the
 233 other loci we identified are in/near genes previously implicated in GWAS of CL/P^{2,14}. For instance,
 234 *PAX7* has been found in multiple GWAS of CL/P including a coding *de novo* variant³⁴. *IRF6*,
 235 *FOXE1* and *NOG* have also been found in multiple GWAS, with additional evidence of functional
 236 common variants found in experimentally validated enhancer regions in these loci^{12,34–36}. Association
 237 of *DLG1* with CL/P has been reported in a recent GWAS³⁷ and experimentally validated in a mouse
 238 model³⁸. Similarly, mouse experiments found cleft and neural tube defects (NTDs) following
 239 alterations of the *SHROOM3* gene³⁸, and *MAFB* has been identified in GWAS of CLP with mouse
 240 models showing its role in palate development²⁹.

The associated SNPs at the 4p13 locus occur within a topologically associating domain containing two genes: *LIMCH1* and *UCHL1*³⁹. This locus was previously found to be suggestively associated with CL/P using GENEVA and POFC subjects¹⁴. To our knowledge, neither of these genes has been directly implicated in development of clefts to date. There is some evidence of altered methylation patterns in peripheral blood for *LIMCH1* found in Han Chinese pedigrees with children affected by NTDs⁴⁰. Although OFCs and NTDs are considered ‘mid-line birth defects’, and supplementing mothers with folate and multivitamins during pregnancy seem to reduce risk to both⁴¹, it remains unknown if the same genes influence their risk. The region around the associated SNPs contains multiple putative craniofacial enhancers derived from epigenomic marks in human fetal craniofacial tissue, and the *LIMCH1* and *UCHL1* genes are decorated with marks associated with active transcription⁴². These data suggest that this locus could play a role in craniofacial development but provide no clues for the opposite effects of SNPs on the risk to CL/P and CP.

Similarly, *SH3PXD2A* was recently implicated in the etiology of CL/P for the first time in a GWAS of individuals from the Netherlands and Belgium³¹. Zebrafish and mouse models support some role of this gene in OFCs³¹. Homozygous disruption of this gene in mutant mice resulted in complete clefts of the secondary palate⁴³. These studies suggest that *SH3PXD2A* might play a role in the pathogenesis of both CL/P and CP, but it is not yet clear how opposite genetic effects of the markers near this gene mechanistically influence risk to subtypes CLP and CP.

We also found opposite effects of associated SNPs for CL and CP near another novel OFC gene, *MIR302F*. There is some evidence that *MIR302* family members regulate *TP63*, a gene found mutated in ectodermal dysplasia-clefting syndrome²⁹. In mouse models, members of the miR-302 family (miR-302 a-d) target different isoforms of the p63 transcription factor, the expression of which is critical for normal lip and palate development. Complete loss of p63 expression leads to CLP in mouse models⁴⁴. Unfortunately, these studies did not include miR-302f, and little is known about the *MIR302F* gene. It is possible that *MIR302F* plays a similar critical role in craniofacial

266 development as the other members of the *MIR302* family.

267 In this manuscript we have annotated the 1p21.3 and 19p13.12 loci with the genes *MIR137HG*
268 and *RAB8A*, respectively. However, these annotations are based on proximity to the most significant
269 SNPs and there is no specific evidence in the literature to support their role in craniofacial development.
270 There are 3 genes in the topological domain containing the associated SNPs of 1p21.3 locus
271 (*MIR137HG*, *MIR2682* and *DPYD*)³⁹, which may be associated with schizophrenia and bipolar
272 disorder⁴⁵. *RAB8A* itself has been shown to be associated with endometrial cancer⁴⁶. It will be an
273 area of future work to replicate and further elucidate our findings near *RAB8A* and *MIR137HG*.

274 Taken together, our study provides strong statistical evidence for possible overlap in the
275 genetic architectures of CL/P and CP. Historically, these two OFC subgroups have been thought
276 to have distinct etiologies based on developmental origins and epidemiologic patterns. Linkage
277 studies first identified significant evidence of linkage for markers in the *FOXE1* region in CL/P
278 multiplex families⁴⁷; subsequent studies confirmed this gene as a risk factor for both CL/P and
279 CP^{11,48}. Linkage analysis identified a susceptibility locus near *TBX22* for CL/P and a later study
280 found mutations in *TBX22* in CP individuals¹⁷. Fine-mapping of translocation breakpoints revealed
281 an important role of *SATB2* in cases with CP; a few years later, a candidate gene study identified
282 significant association of a variant in *SATB2* with CL/P in two Asian populations^{10,17}. There exists
283 some evidence that variants in *IRF6*²⁸, *GRHL3*²⁸, *ARHGAP29*^{49–51} and *MSX1*¹⁷ regions may affect
284 risk to both CL/P and CP in the same direction (often termed as ‘shared genetic risk’). Note, much
285 of this evidence are from patterns of Mendelian inheritance of rare variants in extended pedigrees;
286 genetic overlap of common variants in nonsyndromic OFCs may not follow the same patterns. In
287 recent GWAS, variants near *IRF6*²⁰ and *NOG*^{14,19} have shown weak evidence of decreased risk for
288 one OFC subgroup and increased risk for the other. Among these well-recognized risk genes for
289 OFCs, only *FOXE1* has been successfully implicated as a shared risk gene in GWAS¹². Our method
290 PLACO not only replicated this finding for *FOXE1*, but also provided strong statistical evidence for

genetic overlap at *IRF6* and *NOG*. PLACO found no evidence of genetic overlap between CL/P & CP at common variants (MAF $\geq 5\%$) in/near *SATB2* (chr2:200134004-200316268, including rs6705250 and rs12105015; $p_{\min} = 0.03$), *GRHL3*, *ARHGAP29* or *MSX1* (**Figure S18**). It is possible that rarer variants drive genetic overlap in these regions, and PLACO is currently only applicable to common variants.

In summary, this study advanced our knowledge of the genetic architecture controlling risk to OFCs by enriching the current inventory of OFC-associated genes with novel genes possibly driving common genetic basis of OFC subtypes. Lack of granularity of cleft subtype in animal models precludes experimental validation of our findings. No bioinformatics analysis on existing large-scale databases is equipped to explain how some of these loci could affect the two OFC subgroups in opposing directions. Instead we utilized *in-silico* validation techniques such as statistical simulation experiments, sensitivity analyses, and proof-of-principle analysis. Our extensive *in-silico* validation showed PLACO's robustness to subgroup-specific effects (not a situation of genetic overlap), population-specific differences in MAF, and sample size differences between OFC subgroups. Our proof-of-principle analysis of subtypes CL & CLP using PLACO and replicating those findings with genetic associations of subgroup CL/P emphasized the shared etiology successfully identified by PLACO. More granular functional studies than those currently available are needed to clearly understand the differences in how some of the genes identified here could affect risk of one subtype versus another.

METHODS

The POFC and the GENEVA studies. Case-parent trios ascertained through cases with an isolated, nonsyndromic OFC in the GENEVA study were largely recruited through surgical treatment centers by multiple investigators from Europe (Norway and Denmark), the United States (Iowa, Maryland, Pennsylvania, and Utah) and Asia (People's Republic of China, Taiwan, South Korea,

Singapore, and the Philippines)^{12,49}. Type of cleft, sex, race as well as common environmental risk factors were obtained through direct maternal interview⁴⁹. The research protocol was approved by the Institutional Review Boards (IRB) at the Johns Hopkins Bloomberg School of Public Health and at each participating recruitment site. Written informed consent was obtained from both parents and assent from the child was solicited whenever the child was old enough to understand the purpose of the study.

The POFC study included case-parent trios ascertained through a proband with an isolated nonsyndromic CL/P or CP from multiple populations, and a large number of OFC cases and ethnically matched controls from some of these same populations^{12,52}. We, however, used only the case-parent trios from POFC in this study. Similar information about type of cleft, sex, race and common environmental risk factors were collected through direct maternal interview. The research protocol was approved by the IRB at the University of Pittsburgh and all participating institutions, and informed consent was obtained from all participants.

The distribution of trios by cleft subtype and racial/ethnic groups from both studies is given in **Table S1**. It is to be noted that originally 412 individuals from POFC were included in GENEVA⁵²; we have subsequently removed them from our GENEVA dataset to avoid duplication. Thus, in this article, these two studies represent independent, non-overlapping case-parent trios from three major racial/ethnic groups (European, Asian, and Latin American). Instead of having separate discovery and replication samples, we decided to combine the two studies, which should have improved power to detect genetic associations over a two-stage discovery-replication approach⁵³.

Genotyping, imputation and quality control. Participants in the GENEVA study were genotyped on the Illumina Human610 Quadv1_B array with 589,945 SNPs at the Center for Inherited Disease Research (<https://cidr.jhmi.edu/>). We re-imputed this dataset using the Michigan Imputation Server⁵⁴ to take advantage of more efficient imputation tools and more recent, larger reference panels.

Before imputing, we dropped SNPs with $MAF < 1\%$, performed trio-aware phasing of the haplotypes from the observed genotypes using SHAPEIT2⁵⁵, and original genotyped SNPs on build hg18 were lifted over to hg19 (<http://github.com/sritchie73/liftOverPlink>). We used the 1000 Genomes Phase 3 release 5 reference panel for imputation. Note, trio-aware phasing before imputation is critical; ignoring the family information may lead to biased downstream results⁵⁶. ‘Hard’ genotype calls were made by setting threshold 0.1 within PLINK 2.0 (www.cog-genomics.org/plink/2.0/)⁵⁷. If the calls have uncertainty > 0.1 (i.e., genotype likelihoods < 0.9), they were treated as missing; the rest were regarded as hard genotype calls. We took the following quality control measures: all genotyped SNPs with missingness $> 5\%$, Mendelian error rate $> 5\%$, all SNPs with $MAF < 5\%$ as well as those showing deviation from Hardy-Weinberg equilibrium (HWE) at $p < 10^{-4}$ among parents were dropped. All imputed SNPs were filtered to exclude any with $R^2 < 0.3$ using BCFtools-v1.9 (<https://samtools.github.io/bcftools>). Additionally, individuals with low genotype information or evidence of low-quality DNA, individuals with SNP missingness $> 10\%$, and individuals duplicated across the POFC and GENEVA datasets were excluded. Only complete trios were kept for the final analysis. The final GENEVA dataset contained 6,762,077 autosomal SNPs, including both observed and imputed SNPs having $MAF \geq 5\%$ among parents, for 1,939 complete case-parent trios.

The case-parent trios from the POFC study were genotyped on 539,473 SNPs on the Illumina HumanCore + Exome array. For imputation, data were phased with SHAPEIT2 and imputed using IMPUTE2⁵⁸ to the 1000 Genomes Phase 3 reference panel as described previously⁵². Incomplete trios, trios with parents from different racial/ethnic groups and racial/ethnic groups with insufficient sample sizes for effective imputation were dropped. The same quality control measures as GENEVA were used to remove rare and poor quality SNPs. The final POFC dataset contained 6,350,243 autosomal SNPs, including both observed and imputed SNPs having $MAF \geq 5\%$ among parents, for 1,443 complete case-parent trios.

We meta-analyzed the POFC and the GENEVA studies (which are independent) to increase sample size and power. All SNPs with mismatch in allele or base pair information (hg19) between the two studies were removed. The final meta-analyzed dataset contained 6,761,961 SNPs, which includes all SNPs common to the two studies and SNPs unique to any one study.

PLACO: pleiotropic analysis under a composite null hypothesis. Consider genome-wide studies of two disorders Y_1 and Y_2 based on n_1 and n_2 case-parent trios respectively who were genotyped/imputed or sequenced at p SNPs. For a given SNP, assume an additive genetic model where the relative risks associated with the two disorders are RR_1 and RR_2 . The corresponding genetic effect parameters are respectively $\beta_1 = \log(RR_1)$ and $\beta_2 = \log(RR_2)$. One may assume any other genetic inheritance model, and the following is still applicable. In each study, the genotypic transmission disequilibrium test (gTDT) using a conditional logistic framework^{59,60} may be used to obtain the maximum likelihood estimates $\hat{\beta}_k$ and its standard error \hat{se}_k , which are used to construct the summary statistic $Z_k = \frac{\hat{\beta}_k}{\hat{se}_k}$ for $k = 1, 2$. Since TDTs for trios protect against confounding due to population stratification, one may combine multi-ethnic case-parent trios when analyzing the two disorders.

To implement PLACO, we start with the summary statistics Z_1 and Z_2 from the two disorders across all SNPs genome-wide. For practical purposes, we assume the datasets for the two disorders are independent since case-parent trios are ascertained based on the disease status of the child and it is unlikely to have subjects shared between the two datasets. While the usual multi-trait methods^{23,24} test the null hypothesis of no association of a given SNP with any disorder (i.e., $\beta_1 = 0 = \beta_2$) against the alternative hypothesis that at least one disorder is associated, PLACO tests the composite null hypothesis that at most one disorder is associated with the SNP against the alternative that both disorders are associated²⁷. Mathematically, PLACO tests

$$H_0 : \beta_1\beta_2 = 0 \text{ versus } H_a : \beta_1\beta_2 \neq 0$$

so that rejection of the null hypothesis statistically indicates genetic overlap between disorders. The null hypothesis H_0 is a composite of the global null $\{\beta_1 = 0 = \beta_2\}$, and the sub-nulls $\{\beta_1 = 0, \beta_2 \neq 0\}$ and $\{\beta_1 \neq 0, \beta_2 = 0\}$. Suppose, across the genome, the global null holds with probability π_{00} under which the summary statistics Z_1 and Z_2 have asymptotic standard normal distributions. Further assume that the first sub-null holds with probability π_{01} where Z_1 has a standard normal distribution and Z_2 has a shifted normal distribution $N(\mu_2, 1)$. For a given disorder, the relative risks of SNPs with a non-null effect vary genome-wide. Consequently, there is no fixed value that the mean parameter μ_2 takes, and to capture this variability in effect sizes we assume a random effect on the mean – a $N(0, \tau_2^2)$ distribution. Similarly, assume that the second sub-null holds with probability π_{02} and $Z_2 \sim N(0, 1)$ while Z_1 given μ_1 has a $N(\mu_1, 1)$ distribution, where μ_1 is assumed to follow a $N(0, \tau_1^2)$ distribution.

Thus, under the composite null hypothesis H_0 , PLACO assumes (a) Z_1 and Z_2 are independent $N(0, 1)$ variables when $\{\beta_1 = 0, \beta_2 = 0\}$ holds; (b) Z_1 and Z_2 are independent $N(0, 1)$ and $N(0, 1 + \tau_2^2)$ variables respectively when $\{\beta_1 = 0, \beta_2 \neq 0\}$ holds; and (c) Z_1 and Z_2 are independent $N(0, 1 + \tau_1^2)$ and $N(0, 1)$ variables respectively when $\{\beta_1 \neq 0, \beta_2 = 0\}$ holds. We have described the rationale and other considerations behind this choice of PLACO model previously²⁷. The PLACO test statistic is

$$T_{\text{PLACO}} = Z_1 Z_2$$

and its approximate, asymptotic p-value is given by

$$p_{\text{PLACO}} = \mathbb{P}\left(z_1 z_2 / \sqrt{\text{Var}(Z_1)}\right) + \mathbb{P}\left(z_1 z_2 / \sqrt{\text{Var}(Z_2)}\right) - \mathbb{P}(z_1 z_2)$$

where z_1 and z_2 are the observed Z -scores for the two disorders at any given SNP; $\mathbb{P}(u) = 2 \int_{|u|}^{\infty} \mathbb{f}(x) dx$ is the two-sided tail probability of a normal product distribution at value u ; and $\text{Var}(Z_1)$ and $\text{Var}(Z_2)$ are the estimated marginal variances of these Z -scores under the above distributional model²⁷. Open-source implementation of PLACO in R⁶¹ is available at <https://>

394 [//github.com/RayDebashree/PLACO](https://github.com/RayDebashree/PLACO). While PLACO was originally proposed for two
395 traits from population-based studies (e.g. case-control traits)²⁷, here we showed PLACO can very
396 well be used for family studies (**Supplementary S2**) as long as the summary statistics are obtained
397 after appropriately accounting for all confounding effects, including relatedness and population
398 stratification.

399 **Statistical analyses.** For all analyses presented here, we focused on bi-allelic SNPs with $MAF \geq 5\%$,
400 where MAF is calculated based on only the parents (founders) using PLINK 2.0. For each study
401 separately, we obtained summary statistics of genetic association between each variant and each
402 OFC subgroup using the gTDT model for case-parent trios under an additive genetic model as
403 implemented in R package `trio`⁵⁹ (v3.20.0). We used the gTDT over the allelic TDT because of
404 its several advantages⁶²: gTDT can be more powerful; yields parameter estimates, standard errors
405 along with p-values; and enables direct assessment of RR. However, unlike the allelic TDT⁶³,
406 the gTDT assumes a specific mode of inheritance; here we chose an additive model. The gTDT
407 in `trio` package uses the minor allele of the input dataset as the effect allele. Since the gTDT is
408 applied on each subgroup and each study separately, it is possible that a minor allele is not the same
409 across subgroups and across studies. Therefore, we set the minor allele from the CL/P subgroup
410 in POFC as the effect allele for all analyses. We, then, meta-analyzed the gTDT results over the
411 two studies using inverse-variance weighted fixed effects meta-analysis. We implemented PLACO
412 on the meta-analyzed gTDT results from subgroups CL/P and CP to identify possible genetic
413 overlap between them. To identify regions of significant genetic overlap, we used the conventional
414 genome-wide threshold 5×10^{-8} and also a suggestive threshold of 10^{-6} .

415 We explored if any identified region of genetic overlap was modified by sex. To do this,
416 we first obtained summary statistics from the 1 df SNP×Sex analysis using the gene-environment
417 gTDT model in `trio` package (again assuming additive genetic model); then meta-analyzed the

418 SNP×Sex estimates across the two studies; and finally applied PLACO on the meta-analyzed 1 df
419 SNP×Sex summary statistics. For each analysis, we created Manhattan plots to show signals, and
420 QQ plots to check for potential bias in the association results. For the QQ plots, we also calculated
421 the genomic inflation factors at the 50th percentile ($\lambda_{0.5}$) to quantify the extent of the bulk inflation,
422 and at the 1/10th of a percentile ($\lambda_{0.001}$) to quantify inflation towards the meaningful tail of the
423 distribution. We took the p-values from a given method (e.g., gTDT, PLACO), mapped them to 1
424 df χ^2 statistics, and calculated λ_x as the ratio of empirical 100(1 - x)th percentile of these statistics
425 and the theoretical 100(1 - x)th percentile of 1 df χ^2 distribution.

426 **Stratified analyses.** We considered two stratified analyses: one stratified by racial/ethnic group
427 and the other by cleft subtype (CL, CLP, CP). As described before, genome-wide meta-analyzed
428 gTDT summary statistics were obtained for CL/P and for CP within each of the three major
429 racial/ethnic groups: European, Asian and Latin American. PLACO was applied on CL/P and CP
430 summary data within each racial/ethnic group. For the OFC subtype stratified analyses, meta-analyzed
431 gTDT summary statistics were obtained for each cleft subtype, and then PLACO was applied on
432 each of the three pairwise combinations to compare and contrast results against those from the
433 main analysis.

434 **Locus annotation and candidate gene prioritization.** For each analysis of genetic overlap, we
435 defined independent loci by clumping all the suggestively significant SNPs ($p_{\text{PLACO}} < 10^{-6}$) in
436 a ± 500 Kb radius and with linkage disequilibrium (LD) $r^2 > 0.2$ into a single genetic locus.
437 This clumping was done using FUMA⁶⁴ (SNP2GENE function, v1.3.5e). Since we performed
438 multi-ethnic analysis, we separately used 1000G Phase 3 EUR, EAS and AMR as reference populations
439 for LD calculation. The number of independent loci and the index SNP for each locus (chosen to
440 be the most significant SNP) were the same regardless of the racial/ethnic groups assumed for LD
441 calculation. To define the bounds of each locus, as used in the regional association plots annotated

by effect size directions (**Figures S6, S16 and S17 e**), we took the minimum of lower bounds and the maximum of upper bounds across racial/ethnic groups. We mapped each locus to the gene nearest to the lead SNP using FUMA. We used LocusZoom⁶⁵ to get regional association plots with gene tracks that allowed us to examine detailed evidence of association at each identified locus. For these LocusZoom plots, we used genome build hg19 with no specified LD reference panel due to the multi-ethnic nature of our analysis. The LocusZoom plots for the stratified analyses of the three racial/ethnic groups, however, use the corresponding LD reference panel.

Validation. As mentioned before, we do not have separate discovery and validation samples; combining samples improves power to detect genetic associations over a two-stage discovery-replication approach⁵³. Experimental validation in animal models is not possible due to the lack of granularity of cleft subtypes in mouse or zebrafish models. No current bioinformatics analysis can fully explain the opposite effects of the loci discovered here (where the effect allele increases risk for one OFC subgroup but decreases risk for another). To provide confidence on PLACO's findings, we undertook three complementary approaches: (1) a proof-of-principle analysis of subtypes CL & CLP using PLACO and matching those findings with genetic associations of subgroup CL/P to emphasize the shared etiology that PLACO successfully identified; (2) an *in-silico* validation of PLACO using simulated data; and (3) an assessment of our findings based on existing literature. In particular, our extensive empirical validation involves showing (i) PLACO's robustness to subgroup-specific effects, population-specific differences in MAF, and sample size differences between OFC subgroups; and (ii) massive power gains achieved by PLACO in detecting genetic overlap (whether shared or in opposing directions) compared to other commonly-used variant-level approaches.

***In-silico* evaluation of PLACO using simulated data for OFC trios.** We simulated two bi-ethnic case-parent trio datasets with a total of 2,400 trios mimicking independent studies of CL/P and

CP. We assumed, without loss of generality, that the two ethnic groups have equal sample sizes for a particular OFC subgroup, and considered situations where the OFC subgroups either have comparable (1:1) or unbalanced (3:1) or largely unbalanced (7:1) sample sizes similar to what we saw for the POFC and the GENEVA studies. We simulated the two ethnic groups such that they are different in terms of OFC subgroup prevalence, and in terms of MAF at any given variant. We compared type I error and power of PLACO with the pooled method and the naive approach of declaring genetic overlap when a variant reaches genome-wide significance for the subgroup with the larger sample size and reaches a more liberal significance threshold for the second subgroup. See **Supplementary S2** for more details on our simulation experiments.

Data availability. The POFC and the GENEVA studies are publicly available on dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>, study accession numbers phs000774.v1.p1 and phs000094.v1.p1).

Genome build. All genomic coordinates are given in NCBI Build 37/UCSC hg19.

WEB RESOURCES The R source code for the genetic overlap test PLACO can be found in GitHub (<https://github.com/RayDebashree/PLACO/>). R `trio` package can be found in Bioconductor (<https://bioconductor.org/packages/release/bioc/html/trio.html>).

SUPPLEMENTAL DATA Supplementary information is available on the journal website.

ACKNOWLEDGMENTS This research was supported in part by the NIH grants R03DE029254 (D.R., S.V., J.B.H., T.H.B.), R03DE027121 (S.V., W.Z., M.A.T., T.H.B.), R00DE025060 (E.J.L.), R01DE016148 (M.L.M.) and U24OD023382 (D.R.). It was carried out using computing cluster - the Joint High Performance Computing Exchange - at the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

AUTHOR CONTRIBUTIONS D.R., M.A.T. and T.H.B. designed this study and performed overall project management. W.Z. and J.B.H. performed imputation and processing of the GENEVA data. D.R., S.V.

and W.Z. analyzed the POFC and the GENEVA data. D.R., E.J.L., I.R., M.A.T. and T.H.B. interpreted the scientific findings. D.R. and S.V. performed statistical simulation experiments. M.L.M. and T.H.B. provided the POFC and the GENEVA data respectively. D.R. drafted the initial manuscript. All authors reviewed and approved the final manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

CORRESPONDENCE Correspondence should be addressed to D.R. (email: dray@jhu.edu) or T.H.B. (email: tbeaty1@jhu.edu).

REFERENCES

1. Wehby, G. & Cassell, C. H. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis* **16**, 3–10 (2010).
2. Beaty, T. H., Marazita, M. L. & Leslie, E. J. Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Research* **5** (2016).
3. Christensen, K. & Mortensen, P. B. Facial clefting and psychiatric diseases: a follow-up of the Danish 1936–1987 Facial Cleft cohort. *Cleft Palate Craniofac J* **39**, 392–396 (2002).
4. Leslie, E. J. & Marazita, M. L. Genetics of cleft lip and cleft palate. *Am J Med Genet Part C Semin Med Genet* **163C**, 246–258 (2013).
5. Christensen, K., Juel, K., Herskind, A. M. & Murray, J. C. Long term follow up study of survival associated with cleft lip and palate at birth. *BMJ* **328**, 1405 (2004).
6. Sperber, G. H. in *Cleft lip and palate: from origin to treatment* (ed Wyszynski, D. F.) 5–24 (Oxford University Press, Oxford, 2002).
7. Sivertsen, Å. *et al.* Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *BMJ* **336**, 432–434 (2008).

8. Mangold, E., Ludwig, K. U. & Nöthen, M. M. Breakthroughs in the genetics of orofacial clefting. *Trends Mol Med* **17**, 725–733 (2011).
9. Kondo, S. *et al.* Mutations in *IRF6* cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* **32**, 285–289 (2002).
10. Beaty, T. *et al.* Analysis of candidate genes on chromosome 2 in oral cleft case-parent trios from three populations. *Hum Genet* **120**, 501–518 (2006).
11. Moreno, L. M. *et al.* *FOXE1* association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. *Hum Mol Genet* **18**, 4879–4896 (2009).
12. Leslie, E. J. *et al.* Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between *FOXE1* and all orofacial clefts, and *TP63* and cleft lip with or without cleft palate. *Hum Genet* **136**, 275–286 (2017).
13. Ludwig, K. U. *et al.* Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip \pm cleft palate and cleft palate only. *Hum Mol Genet* **26**, 829–842 (2017).
14. Carlson, J. C. *et al.* A systematic genetic analysis and visualization of phenotypic heterogeneity among orofacial cleft GWAS signals. *Genet Epidemiol* **43**, 704–716 (2019).
15. He, M. *et al.* Genome-wide analyses identify a novel risk locus for nonsyndromic cleft palate. *J Dent Res* **10**, 1–8 (2020).
16. Marazita, M. & Leslie, E. in *Comprehensive cleft care* (eds Losee, J. & Kirschner, R.) 207–224 (CRC Press, Boca Raton FL, 2016).
17. Rahimov, F., Jugessur, A. & Murray, J. C. Genetics of nonsyndromic orofacial clefts. *Cleft Palate Craniofac J* **49**, 73–91 (2012).
18. Carlson, J. C. *et al.* Identifying genetic sources of phenotypic heterogeneity in orofacial clefts by targeted sequencing. *Birth Defects Res* **109**, 1030–1038 (2017).

19. Moreno Uribe, L. *et al.* A population-based study of effects of genetic loci on orofacial clefts. *J Dent Res* **96**, 1322–1329 (2017).
20. Huang, L. *et al.* Genetic factors define CPO and CLO subtypes of nonsyndromic orofacial cleft. *PLoS Genet* **15**, e1008357 (2019).
21. Howe, L. J. *et al.* Investigating the shared genetics of non-syndromic cleft lip/palate and facial morphology. *PLoS Genet* **14**, e1007501 (2018).
22. Carlson, J. C. *Methods for family-based designs in genetic epidemiology studies* PhD thesis (University of Pittsburgh, 2017), 44–44.
23. Ray, D. & Chatterjee, N. Effect of non-normality and low count variants on cross-phenotype association tests in GWAS. *Eur J Hum Genet* **28**, 300–312 (2020).
24. Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol* **7**, 170125 (2017).
25. Fischer, S. T., Jiang, Y., Broadaway, K. A., Conneely, K. N. & Epstein, M. P. Powerful and robust cross-phenotype association test for case-parent trios. *Genet Epidemiol* **42**, 447–458 (2018).
26. Schaid, D. J. *et al.* Multivariate generalized linear model for genetic pleiotropy. *Biostatistics* **20**, 111–128 (2019).
27. Ray, D. & Chatterjee, N. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *bioRxiv* (to appear in *PLOS Genet*), <https://doi.org/10.1101/2020.04.11.037630> (2020).
28. Schutte, B. C., Saal, H. M., Goudy, S. & Leslie, E. in *GeneReviews* (University of Washington, Seattle, 2014).
29. Dixon, M. J., Marazita, M. L., Beaty, T. H. & Murray, J. C. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* **12**, 167–178 (2011).

30. Yu, Y. *et al.* Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat Commun* **8**, 1–11 (2017).
31. Van Rooij, I. A. *et al.* Non-syndromic cleft lip with or without cleft palate: genome-wide association study in europeans identifies a suggestive risk locus at 16p12.1 and supports *SH3PXD2A* as a clefting susceptibility gene. *Genes* **10**, 1023 (2019).
32. Khrantsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat Rev Genet* **20**, 173–190 (2019).
33. Harville, E. W., Wilcox, A. J., Lie, R. T., Vindenes, H. & Åbyholm, F. Cleft lip and palate versus cleft lip only: are they distinct defects? *Am J Epidemiol* **162**, 448–453 (2005).
34. Leslie, E. J. *et al.* Identification of functional variants for cleft lip with or without cleft palate in or near *PAX7*, *FGFR2*, and *NOG* by targeted sequencing of GWAS loci. *Am J Hum Genet* **96**, 397–411 (2015).
35. Rahimov, F. *et al.* Disruption of an AP-2 α binding site in an *IRF6* enhancer is associated with cleft lip. *Nat Genet* **40**, 1341–1347 (2008).
36. Sylvester, B. *et al.* A synonymous exonic splice silencer variant in *IRF6* as a novel and cryptic cause of non-syndromic cleft lip and palate. *Genes* **11**, 903 (2020).
37. Mostowska, A. *et al.* Common variants in *DLG1* locus are associated with non-syndromic cleft lip with or without cleft palate. *Clin Genet* **93**, 784–793 (2018).
38. Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A. & Richardson, J. E. Mouse genome database (mgd) 2019. *Nucleic Acids Res* **47**, D801–D806 (2019).
39. Wang, Y. *et al.* The 3D genome browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* **19**, 1–12 (2018).
40. Zhang, R. *et al.* A unique methylation pattern co-segregates with neural tube defect statuses in Han Chinese pedigrees. *Neurol Sci* **38**, 2153–2164 (2017).

41. Wilson, R. D. *et al.* Pre-conception folic acid and multivitamin supplementation for the primary and secondary prevention of neural tube defects and other folic acid-sensitive congenital anomalies. *J Obstet Gynaecol Can* **37**, 534–549 (2015).
42. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J. P. & Cotney, J. High-resolution epigenomic atlas of human embryonic craniofacial development. *Cell Rep* **23**, 1581–1597 (2018).
43. Cejudo-Martin, P. *et al.* Genetic disruption of the *sh3pxd2a* gene reveals an essential role in mouse development and the existence of a novel isoform of tks5. *PLoS One* **9**, e107674 (2014).
44. Warner, D. R. *et al.* Micro RNA expression profiling of the developing murine upper lip. *Dev Growth Differ* **56**, 434–447 (2014).
45. Duan, J. *et al.* A rare functional noncoding variant at the GWAS-implicated *MIR137/MIR2682* locus might confer risk to schizophrenia and bipolar disorder. *Am J Hum Genet* **95**, 744–753 (2014).
46. Bie, Y. & Zhang, Z. *RAB8A* a new biomarker for endometrial cancer? *World J Surg Oncol* **12**, 371 (2014).
47. Marazita, M. L. *et al.* Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Hum Hered* **68**, 151–170 (2009).
48. Ludwig, K. *et al.* Strong association of variants around *FOXE1* and orofacial clefting. *J Dent Res* **93**, 376–381 (2014).
49. Beaty, T. H. *et al.* A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near *MAFB* and *ABCA4*. *Nat Genet* **42**, 525–529 (2010).

50. Liu, H. *et al.* Identification of common non-coding variants at 1p22 that are functional for non-syndromic orofacial clefting. *Nat Commun* **8**, 14759 (2017).
51. Liu, H. *et al.* Exome sequencing provides additional evidence for the involvement of *ARHGAP29* in Mendelian orofacial clefting and extends the phenotypic spectrum to isolated cleft palate. *Birth Defects Res* **109**, 27–37 (2017).
52. Leslie, E. J. *et al.* A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24. 2, 17q23 and 19q13. *Hum Mol Genet* **25**, 2862–2872 (2016).
53. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38**, 209–213 (2006).
54. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
55. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am J Hum Genet* **93**, 687–696 (2013).
56. Taub, M. A., Schwender, H., Beaty, T. H., Louis, T. A. & Ruczinski, I. Incorporating genotype uncertainties into the genotypic TDT for main effects and gene-environment interactions. *Genet Epidemiol* **36**, 225–234 (2012).
57. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** (2015).
58. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

59. Schwender, H., Taub, M. A., Beaty, T. H., Marazita, M. L. & Ruczinski, I. Rapid testing of SNPs and gene-environment interactions in case-parent trio data based on exact analytic parameter estimation. *Biometrics* **68**, 766–773 (2012).
60. Schwender, H. *et al.* Detecting disease variants in case-parent trio studies using the bioconductor software package `trio`. *Genet Epidemiol* **38**, 516–522 (2014).
61. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2018). <https://www.R-project.org/>.
62. Fallin, D., Beaty, T., Liang, K. Y. & Chen, W. Power comparisons for genotypic vs. allelic TDT methods with >2 alleles. *Genet Epidemiol* **23**, 458–461 (2002).
63. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**, 506–516 (1993).
64. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
65. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

Table 1: Association results for the most significant markers from the 9 loci showing statistical evidence of genetic overlap between CL/P and CP, along with 3 additional loci of genetic overlap between pairwise OFC subtypes. These loci were identified by PLACO at a suggestive threshold of 10^{-6} . The genetic overlap analysis is based on all trios from both POFC and GENEVA for CL/P & CP, or CL & CP, or CLP & CP, or CL & CLP. The different types of novel genes are marked by * or ‡. The “No. of trios” columns give the numbers of complete informative case-parent trios as used by the gTDT method in analyzing each OFC subgroup/subtype.

(Caption in the previous page.)

Locus	Nearest gene	rsID	Position (hg19)	Effect allele (freq.)	CL/P			CP			CL/P & CP
					gTDT p-value	gTDT RR	No. of trios	gTDT p-value	gTDT RR	No. of trios	PLACO p-value
1p36.13	<i>PAX7</i> [‡]	rs1339063	18989575	T (0.33)	2.9×10^{-8}	1.27	1709	5.6×10^{-3}	0.77	385	6.9×10^{-8}
1q32.2	<i>IRF6</i>	rs72741048	209989092	T (0.34)	4.0×10^{-17}	0.70	1799	2.8×10^{-3}	1.30	414	4.3×10^{-12}
3q29	<i>DLG1</i> [‡]	rs12632559	196803647	C (0.42)	1.7×10^{-3}	1.15	1521	2.1×10^{-5}	0.67	355	5.3×10^{-7}
4p13	<i>LIMCH1</i> ^{‡*}	rs9291207	41649103	C (0.3)	6.5×10^{-5}	0.84	1611	8.0×10^{-4}	1.38	347	5.0×10^{-7}
4q21.1	<i>SHROOM3</i> [‡]	rs4422437	77514866	G (0.33)	6.8×10^{-7}	1.23	1841	9.4×10^{-3}	0.80	431	8.1×10^{-7}
9q22.33	<i>FOXE1</i>	rs12347191	100619719	C (0.25)	8.9×10^{-6}	0.81	1526	1.1×10^{-3}	0.73	351	1.7×10^{-7}
17q22	<i>NOG</i>	rs4794658	54766218	T (0.36)	1.9×10^{-6}	1.21	1951	1.1×10^{-3}	0.75	407	6.0×10^{-8}
19p13.12	<i>RAB8A</i> ^{‡*}	rs7252188	16228701	A (0.39)	3.2×10^{-3}	0.89	1844	9.2×10^{-6}	0.69	455	6.8×10^{-7}
20q12	<i>MAFB</i> [‡]	rs6016392	39246610	A (0.5)	8.3×10^{-10}	1.27	2004	3.9×10^{-2}	1.18	452	9.9×10^{-7}
					CL			CP			CL & CP
18q12.1	<i>MIR302F</i> ^{‡*}	rs11083400	28056959	T (0.45)	3.1×10^{-3}	0.79	484	1.6×10^{-5}	1.45	424	6.2×10^{-7}
					CLP			CP			CLP & CP
10q24.33	<i>SH3PXD2A</i> [‡]	rs11191818	105591779	G (0.31)	2.7×10^{-4}	0.82	1102	5.1×10^{-4}	1.40	341	9.2×10^{-7}
					CL			CLP			CL & CLP
1p21.3	<i>MIR137HG</i> ^{‡*}	rs2802532	98528830	C (0.09)	7.3×10^{-5}	0.59	217	3.5×10^{-5}	1.41	562	2.2×10^{-8}

[‡]Genes that have not previously been suggested as regions of genetic overlap between OFC subgroups in linkage or association studies.

*Genes that have not been previously implicated in OFC genetics.

Abbreviations: Chr, chromosome; CL/P, cleft lip with or without palate; CP, cleft palate; Freq., frequency; gTDT, genotypic transmission disequilibrium test; OFC, orofacial cleft; PLACO, pleiotropic analysis under composite null hypothesis; RR, relative risk (with respect to the reported effect allele)

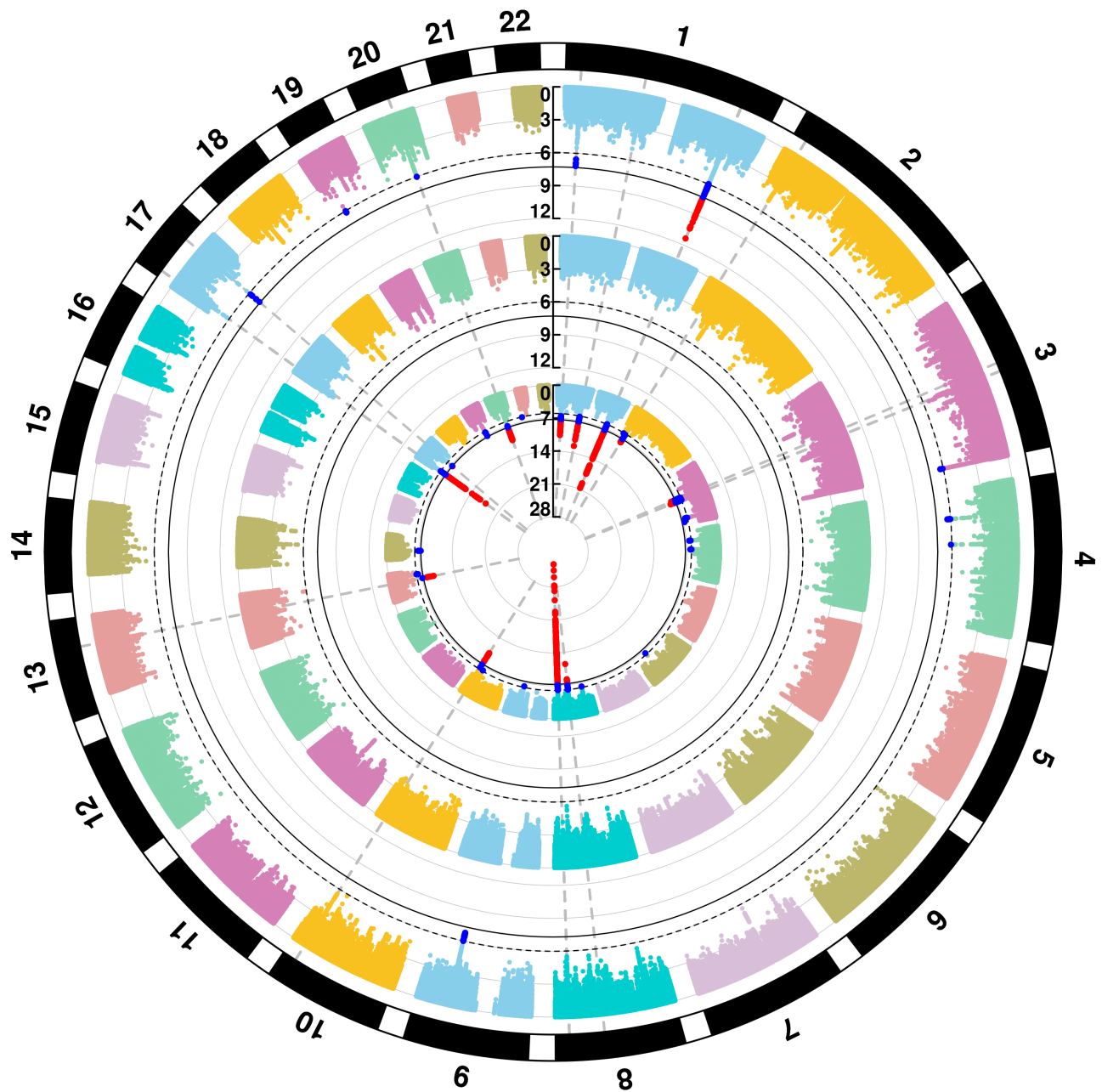
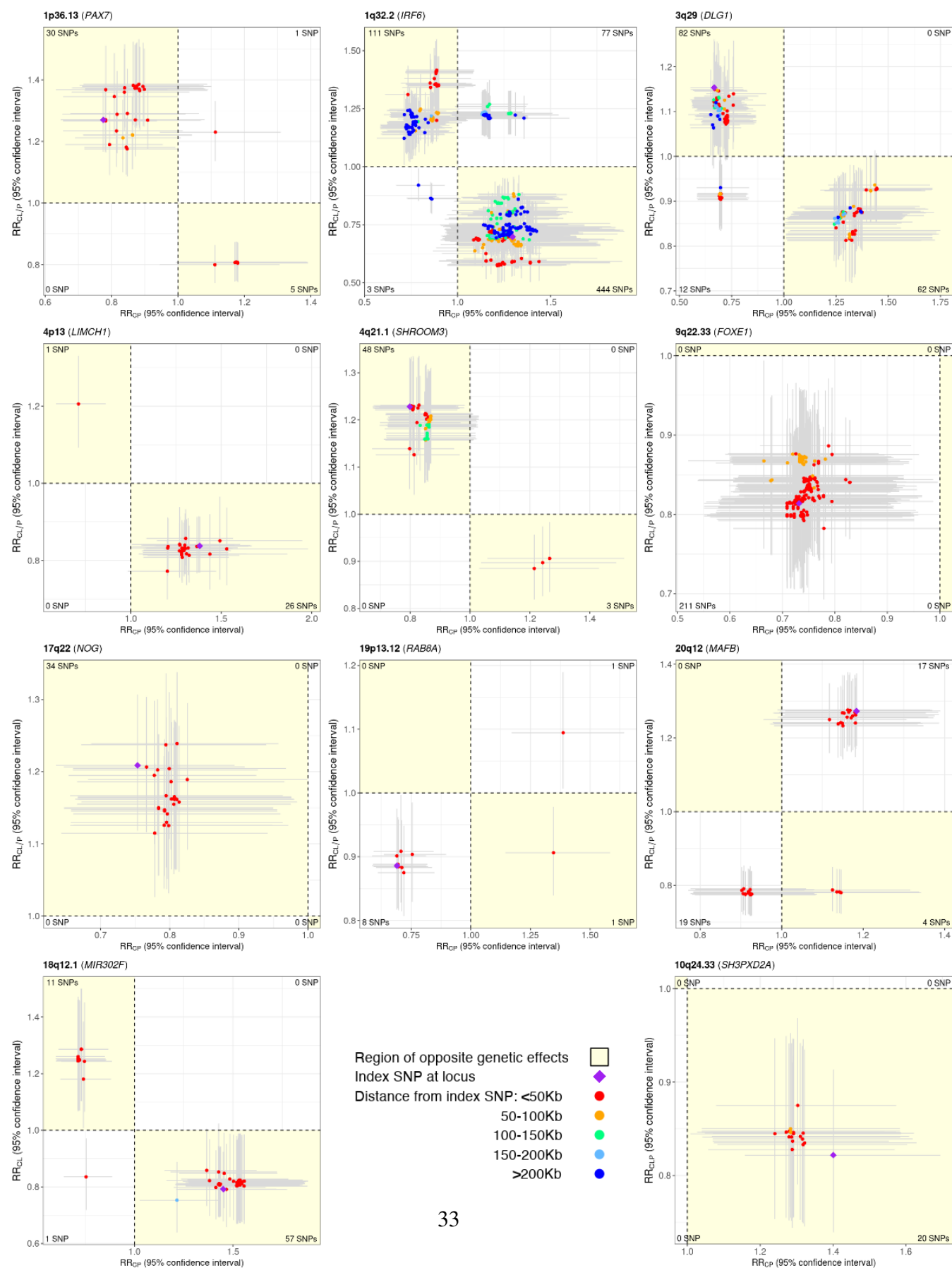


Figure 1: Manhattan plots for genome-wide analyses of OFC subgroups. The plots are of negative log-transformed p-values from the analysis of cleft lip with or without cleft palate (CL/P, innermost circle), of cleft palate (CP, intermediate circle), and of genetic overlap between CL/P & CP using PLACO (outermost circle). The chromosome numbers 1–22 are indicated along the outermost circumference. Solid black and dashed black circular lines are used in all plots to indicate the conventional genome-wide significance threshold 5×10^{-8} and a less stringent suggestive threshold 10^{-6} respectively. The variants exceeding the genome-wide and the liberal thresholds are respectively colored in red and bright blue.



(Caption in the next page.)

Figure 2: Scatter plot of relative risk (RR) estimates, along with corresponding 95% confidence intervals (CIs), for variants in the 9 loci showing statistical evidence of genetic overlap between CL/P & CP, along with 2 additional loci of genetic overlap between component OFC subtypes. RR estimates are color annotated based on distance of SNPs from the index/lead SNP. LD-based color annotation is not used since these RR estimates are from multi-ethnic analyses and consequently, there is no unique LD between SNPs. Horizontal (vertical) error bar around each RR estimate corresponds to the 95% CI for the OFC subgroup represented on the x-axis (y-axis). The region depicting opposite genetic effects of SNPs for the 2 OFC subgroups is shaded in yellow. The number of SNPs in each quadrant is printed in the corresponding corner of the plot. The SNPs plotted here are in ± 500 Kb radius and in LD $r^2 > 0.2$ with the index/lead SNP, and further screened out SNPs with PLACO p-value $> 10^{-3}$ from the respective genetic overlap analysis. These plots show genetically distinct etiology of CL/P and CP at 6 loci (i.e., overlapping genetic etiology with opposite effects), and of the component OFC subtypes at 2 other loci as depicted by the SNPs in the yellow shaded region.

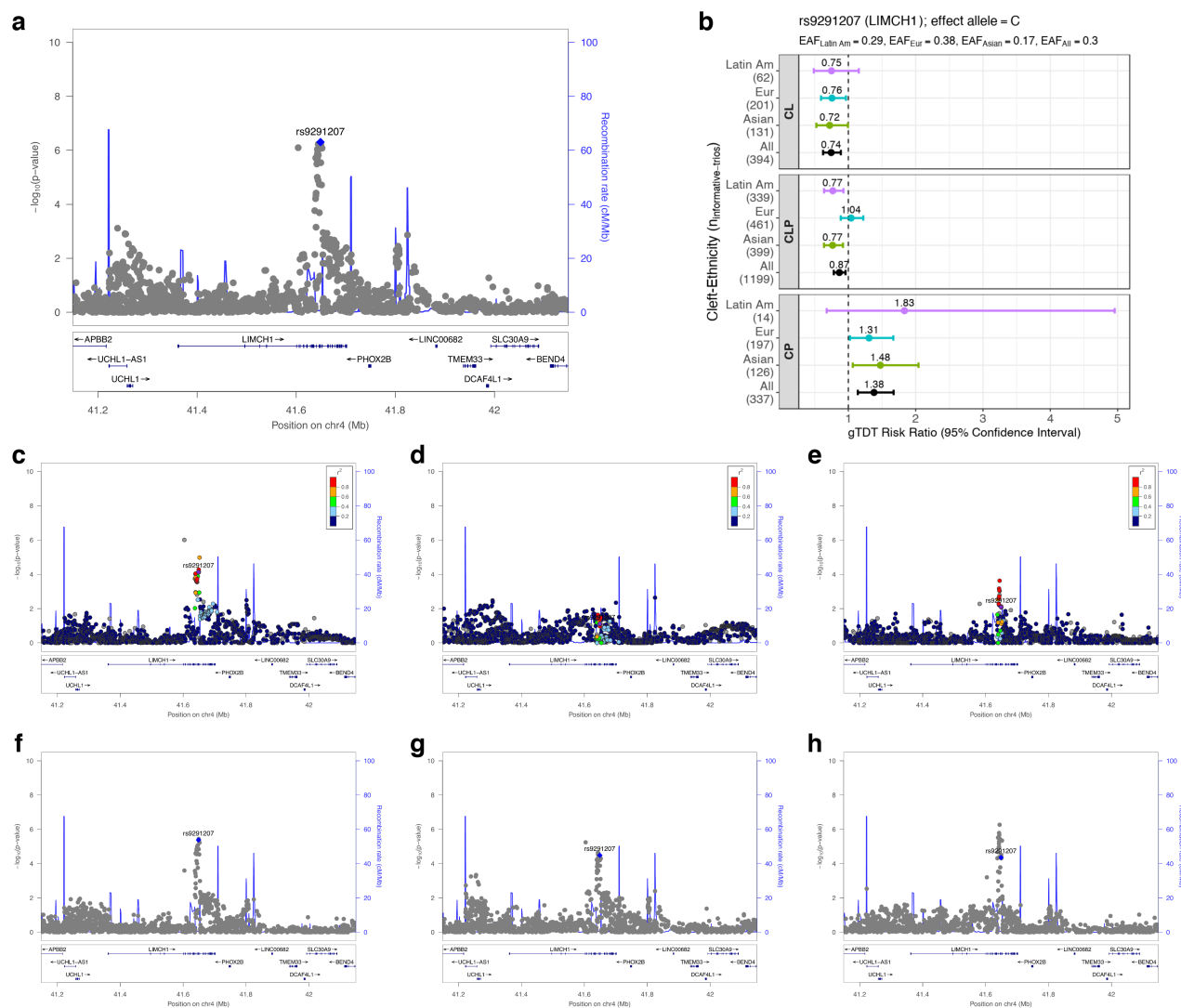


Figure 3: Regional association plots for 4p13 (*LIMCH1*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of (a) CL/P & CP (multi-ethnic), (c) CL/P & CP in Asian ancestry, (d) CL/P & CP in European ancestry, (e) CL/P & CP in Latin American ancestry, (f) CL & CP (multi-ethnic), (g) CLP & CP (multi-ethnic), (h) CL & CLP (multi-ethnic). The blue or purple diamond represents the most strongly associated SNP in the region showing evidence of genetic overlap. For stratified analyses across racial/ethnic groups, the colors of the SNPs represent their LD with the lead SNP (the most strongly associated SNP from panel a), as shown in the color legend. For combined multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used. Panel b shows relative risk estimates of the lead SNP and their 95% confidence intervals as obtained from the gTDT analyses.

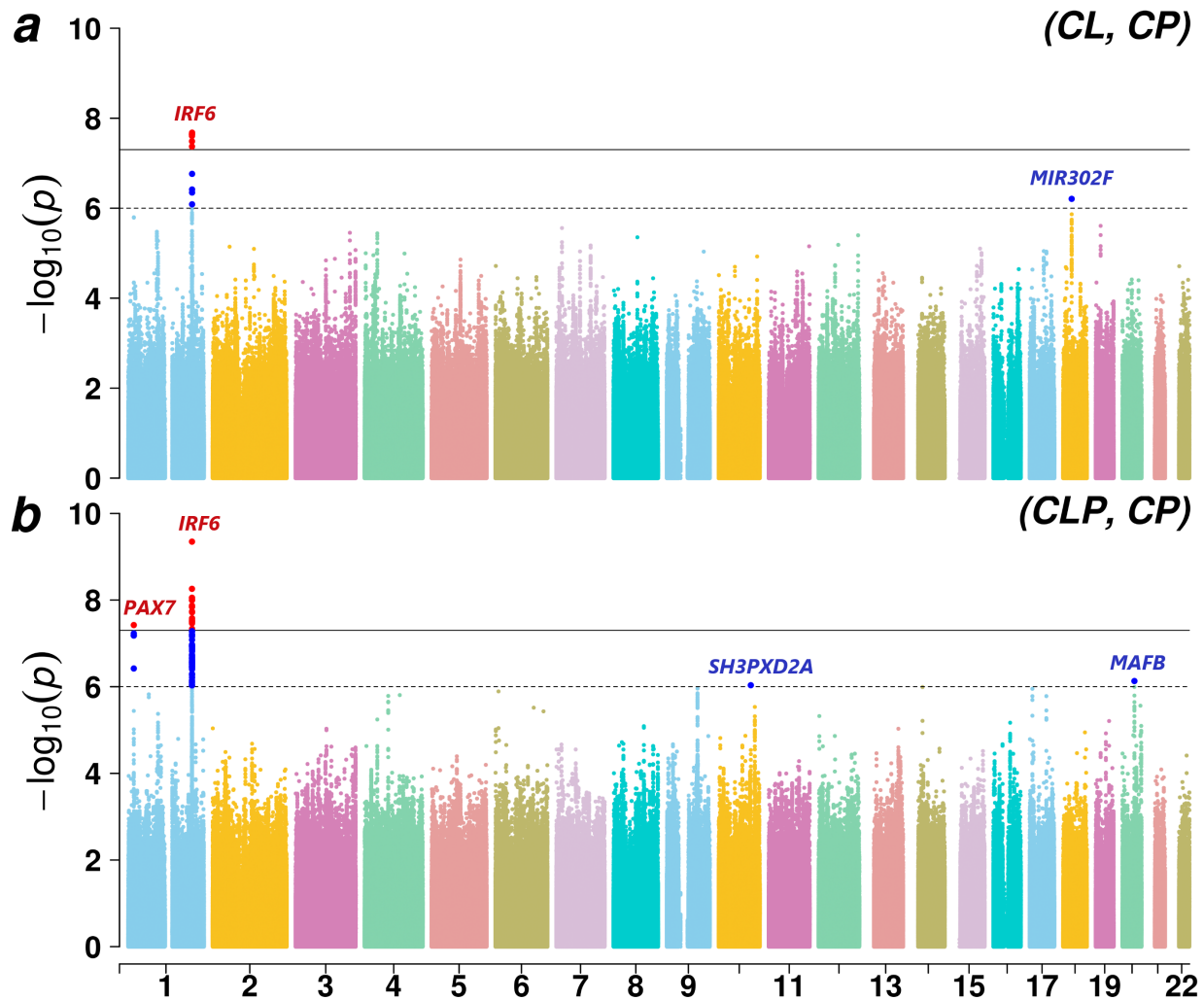


Figure 4: Manhattan plots for genome-wide analyses of genetic overlap between pairs of OFC subtypes: CL & CP, and CLP & CP. The chromosome numbers 1–22 are indicated along the x-axis. Solid black and dashed black lines are used in both plots to indicate the conventional genome-wide significance threshold 5×10^{-8} and a less stringent suggestive threshold 10^{-6} respectively. The variants exceeding the genome-wide and the liberal thresholds are respectively colored in red and bright blue.

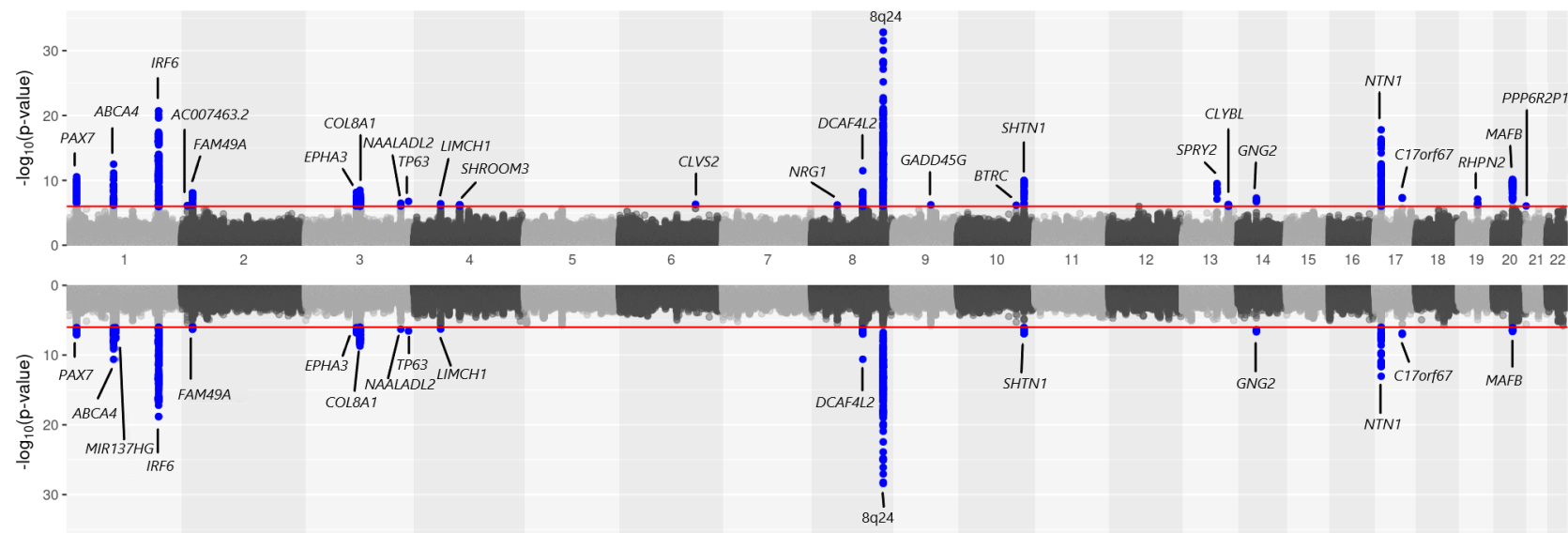


Figure 5: Mirrored Manhattan plot of genome-wide analyses of CL/P, and CL & CLP. Results from genetic associations of CL/P using the gTDT are shown in the upper panel. Results from shared genetic associations between CL & CLP using PLACO are shown in the lower panel. The red horizontal lines in the two panels indicate a suggestive significance threshold of 10^{-6} .