PLOS GENETICS

RESEARCH ARTICLE

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

Debashree Ray[1,2]*, Sowmya Venkataraghavan[1], Wanying Zhang[1], Elizabeth J. Leslie[3], Jacqueline B. Hetmanski[1], Seth M. Weinberg[4,5,6], Jeffrey C. Murray[7], Mary L. Marazita[4,5,6], Ingo Ruczinski[2], Margaret A. Taub[2], Terri H. Beaty[1,2]*

1 Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **3** Department of Human Genetics, School of Medicine, Emory University, Atlanta, Georgia, United States of America, **4** Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **5** Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **6** Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **7** Department of Pediatrics, University of Iowa Carver College of Medicine, Iowa City, Iowa, United States of America

* dray@jhu.edu (DR); tbeaty1@jhu.edu (THB)

OPEN ACCESS

Citation: Ray D, Venkataraghavan S, Zhang W, Leslie EJ, Hetmanski JB, Weinberg SM, et al. (2021) Pleiotropy method reveals genetic overlap between orofacial clefts at multiple novel loci from GWAS of multi-ethnic trios. PLoS Genet 17(7): e1009584. https://doi.org/10.1371/journal.pgen.1009584

Editor: Heather J Cordell, Newcastle University, UNITED KINGDOM

Received: November 23, 2020

Accepted: May 6, 2021

Published: July 9, 2021

Data Availability Statement: The POFC and the GENEVA studies are publicly available on dbGaP (https://www.ncbi.nlm.nih.gov/gap/, study accession numbers phs000774.v2.p1 and phs000094.v1.p1 respectively).

Funding: 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 (S.M.W., M. L.M.) and U24OD023382 (D.R.). The funders did

Abstract

Based on epidemiologic and embryologic patterns, nonsyndromic orofacial clefts– the most common craniofacial birth defects in humans– are commonly categorized into cleft lip with or without cleft palate (CL/P) and cleft palate alone (CP), which are traditionally considered to be etiologically distinct. However, some evidence of shared genetic risk in *IRF6*, *GRHL3* and *ARHGAP29* regions exists; only *FOXE1* has been recognized as significantly associated with both CL/P and CP in genome-wide association studies (GWAS). We used a new statistical approach, PLACO (pleiotropic analysis under composite null), on a combined multi-ethnic GWAS of 2,771 CL/P and 611 CP case-parent trios. At the genome-wide significance threshold of 5×10^{-8}, PLACO identified 1 locus in 1q32.2 (*IRF6*) that appears to increase risk for one OFC subgroup but decrease risk for the other. At a suggestive significance threshold of 10^{-6}, we found 5 more loci with compelling candidate genes having opposite effects on CL/P and CP: 1p36.13 (*PAX7*), 3q29 (*DLG1*), 4p13 (*LIMCH1*), 4q21.1 (*SHROOM3*) and 17q22 (*NOG*). Additionally, we replicated the recognized shared locus 9q22.33 (*FOXE1*), and identified 2 loci in 19p13.12 (*RAB8A*) and 20q12 (*MAFB*) that appear to influence risk of both CL/P and CP in the same direction. We found locus-specific effects may vary by racial/ethnic group at these regions of genetic overlap, and failed to find evidence of sex-specific differences. We confirmed shared etiology of the two OFC subtypes comprising CL/P, and additionally found suggestive evidence of differences in their pathogenesis at 2 loci of genetic overlap. Our novel findings include 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. Our *in-silico* validation showed PLACO is robust to subtype-specific effects, and can achieve massive power gains over existing approaches for identifying genetic overlap between disease subtypes. In summary, we found suggestive evidence for

not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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

Author summary

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 either in the same direction or in opposite directions. 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. Of these loci, 2 were identified at the genome-wide threshold, and the rest at a suggestive significance threshold of 10^{-6}. Locus-specific effects appear to vary by racial/ethnic group at these regions of genetic overlap. 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 suggestive evidence for new genetic regions and confirmed some recognized OFC genes either exerting shared risk or with opposite effects on risk to OFC subtypes.

Introduction

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

Nonsyndromic OFCs typically manifest as a gap in the upper lip ('cleft lip' or CL) or the roof of the mouth ('cleft palate' or CP) or both ('cleft lip and palate' or CLP). Based on epidemiologic evidence, prevalence rates and the embryologic period when they develop, the subtypes CL and CLP are typically grouped together as the subgroup CL/P (cleft lip with or without palate) [6–8], while CP alone forms the other subgroup. CL/P and CP have been historically analyzed separately [9–12]. While genetic studies have identified nearly 40 genetic regions (or loci) as significantly associated with risk to CL/P, fewer, around 10 loci, have been identified for CP [2, 13–15]. The findings for CP have mostly been identified in the Han Chinese population [15]. These genetic studies, ranging from candidate gene approaches to genome-wide studies along with analysis of animal models, have provided insights into the genetic etiology of nonsyndromic OFCs [16]. Together, these genetic regions explain no more than a quarter of the estimated total heritability of risk to OFCs [17].

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

between genetic overlap and genetic heterogeneity. While genetic heterogeneity may refer to shared genetic effects as well as subtype-specific effects (which may mean a non-null effect on one subtype and no effect on the other), genetic overlap refers to non-null genetic effects on both subtypes that may or may not be equal in magnitude and/or direction. The usual approach for identifying genetic overlap in OFCs is to compare the significant findings among common variants from one subtype with those from the other [11, 13, 20, 21]. However, the discovery of the associated variants in the first place may be under-powered in genome-wide association study (GWAS) of each subtype separately. For instance, success of discovery genetics has been elusive for CP, which could partially be due to smaller sample sizes of CP [20] reflecting its lower birth prevalence. Another approach involves testing how well polygenic risk scores for one subtype can explain variation for another [13, 22], which describes overall genetic sharing, does not implicate specific regions of overlap (novel or otherwise), and may indicate lack of overlap when one subtype has a much smaller sample size [13, 14]. One strategy is the 'pooled method' GWAS analysis [12], where all the OFC subtypes are pooled together in a combined analysis of all OFCs. *FOXE1* has been successfully implicated as a shared risk gene using this approach on common single nucleiotide polymorphisms (SNPs), along with bioinformatics analyses of top SNPs that helped prioritize this gene [12]. While association signals from the pooled method may be driven by shared risk variants between subtypes, the pooled method does not necessarily capture only shared signals [23], especially if sample sizes are widely different between the subtypes (e.g., the CL/P group is almost always much larger than the CP group) or if strong genetic effects exist in one group but not the other. Furthermore, if a locus is hypothesized to have opposite genetic effects on the two subtypes (e.g., *NOG* [14, 20]), the pooling technique will dilute any signal and consequently will be under-powered to detect genetic overlap. Use of multi-trait methods in OFC genetic studies, such as the ones commonly used in population-based GWAS of complex traits [24–27], faces multiple obstacles: the disjoint nature of the subtypes (i.e., absence of subjects with both traits); the qualitative (binary) nature of the traits; and/or the case-parent trio design typically used to study multi-ethnic samples with 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 studies [28], here we found it can also help identify genetic variants simultaneously influencing risk in two disease subgroups from family-based studies. 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 on common SNPs 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 gene [29]: 1q32.2 (*IRF6*, $p = 4.3 \times 10^{-12}$).

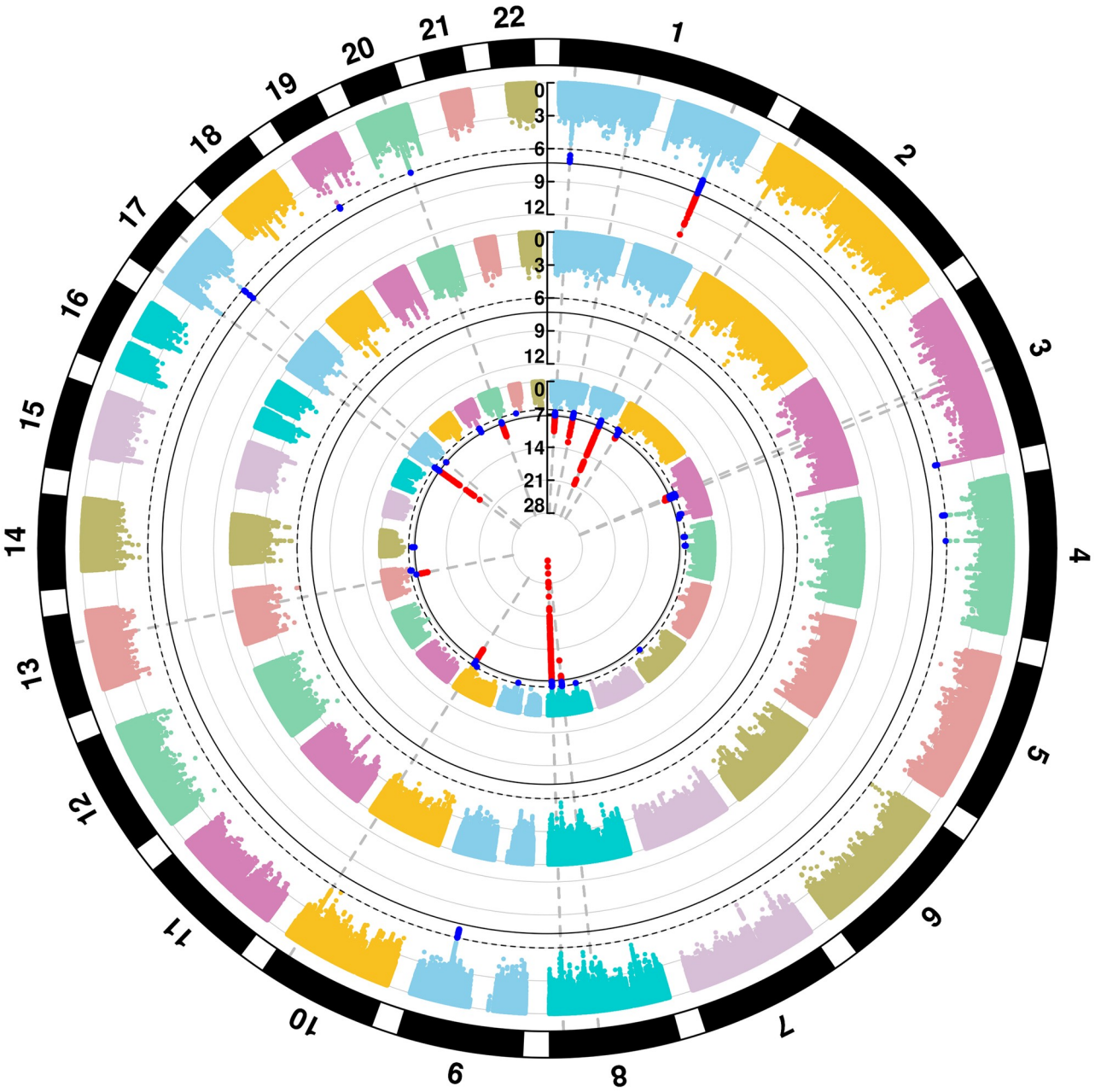

Fig 1. Manhattan plots for genome-wide analyses of OFC subgroups. The plots are of negative log-transformed p-values from the analysis of cleft lip with or without cleft palate (CL/P, innermost circle), of cleft palate (CP, intermediate circle), and of genetic overlap between CL/P & CP using PLACO (outermost 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.

https://doi.org/10.1371/journal.pgen.1009584.g001

We found 2 promising loci in 1p36.13 (*PAX7*, $p = 6.9 \times 10^{-8}$) and 17q22 (*NOG*, $p = 6.0 \times 10^{-8}$), just missing the GWAS significance threshold (Fig 1). Additionally, at a suggestive significance threshold of 10^{-6}, 6 loci showed evidence for genetic overlap between CL/P & CP: 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 (*FOXE1*, $p = 1.7 \times 10^{-7}$), 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 elsewhere [16]. QQ plots from all our analyses show deviation from the null only in the tail end of the distribution of p-values (S1 and S2 Figs), indicating genetic signals rather than any systemic bias.

Six out of 9 loci yielding novel suggestive evidence for genetic overlap between CL/P & CP

Of the 9 loci, genetic overlap at SNPs in/near genes *IRF6* [29], *FOXE1* [12] and *NOG* [14, 20] 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 relative risk (RR) estimates (Table 1). In other words, the effect alleles at the lead SNPs of *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 (Fig 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 (Fig 3B and S3B–S5B Figs). 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 before [12]. Additionally, the *IRF6* region appears to harbor at least 2 distinct loci: one with shared genetic effects, and another with opposite effects (Fig 2 and S6 Fig). Evidence for both shared and opposite effects at *IRF6* has been reported previously [14, 30].

Genetic overlap between subgroups CL/P & CP is consistent with overlap identified between pairwise OFC subtypes CL & CP and CLP & CP

To gain a better understanding of which subtypes are driving this evidence of genetic overlap between CL/P & CP, we applied PLACO on the pairwise component OFC subtypes (Fig 4). The *PAX7*, *SHROOM3*, *FOXE1* and *MAFB* loci seem to be driven by the common genetic basis of subtypes CLP & CP at these loci (S7G–S10G Figs). The rest of the loci (*LIMCH1*, *IRF6*, *DLG1*, *NOG* and *RAB8A*) appear to be driven by genetic overlap between CL & CP as well as CLP & CP (Fig 3F and 3G and S3F and S3G, S4F and S4G, S5F and S5G and S11F and S11G Figs).

Identification of additional novel regions of genetic overlap between component OFC subtypes

PLACO revealed 2 loci in 18q12.1 (*MIR302F*, $p = 6.2 \times 10^{-7}$) and 10q24.33 (*SH3PXD2A*, $p = 9.2 \times 10^{-7}$) associated with CL & CP, and CLP & CP subtypes respectively at a suggestive threshold of 10^{-6} (Table 1). Effect alleles at the lead SNPs of both these loci increase risk for

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. CL/P & CP PLACO p-value $< 5 \times 10^{-8}$ for a SNP indicates statistically significant association of the SNP with both CL/P and CP at the genome-wide level. Similar interpretations for the other PLACO p-values: CL & CP, or CLP & CP, or CL & CLP. PLACO p-value $< 10^{-6}$ is considered suggestive evidence of genetic overlap.

Locus	Nearest gene	rsID	Position (hg19)	Effect allele (freq.)	CL/P			CP			CL/P & CP
					gTDT p-value	gTDT RR [95% CI]	No. of trios	gTDT p-value	gTDT RR [95% CI]	No. of trios	PLACO p-value
1p36.13	*PAX7* ‡	rs1339063	18989575	T (0.33)	2.9×10^{-8}	1.27 [1.17, 1.38]	1709	5.6×10^{-3}	0.77 [0.65, 0.93]	385	6.9×10^{-8}
1q32.2	*IRF6*	rs72741048	209989092	T (0.34)	4.0×10^{-17}	0.70 [0.64, 0.76]	1799	2.8×10^{-3}	1.30 [1.09, 1.54]	414	4.3×10^{-12}
3q29	*DLG1* ‡	rs12632559	196803647	C (0.42)	1.7×10^{-3}	1.15 [1.05, 1.26]	1521	2.1×10^{-5}	0.67 [0.55, 0.80]	355	5.3×10^{-7}
4p13	*LIMCH1* ‡*	rs9291207	41649103	C (0.3)	6.5×10^{-5}	0.84 [0.77, 0.91]	1611	8.0×10^{-4}	1.38 [1.14, 1.67]	347	5.0×10^{-7}
4q21.1	*SHROOM3* ‡	rs4422437	77514866	G (0.33)	6.8×10^{-7}	1.23 [1.13, 1.33]	1841	9.4×10^{-3}	0.80 [0.67, 0.95]	431	8.1×10^{-7}
9q22.33	*FOXE1*	rs12347191	100619719	C (0.25)	8.9×10^{-6}	0.81 [0.74, 0.89]	1526	1.1×10^{-3}	0.73 [0.61, 0.88]	351	1.7×10^{-7}
17q22	*NOG*	rs4794658	54766218	T (0.36)	1.9×10^{-6}	1.21 [1.12, 1.31]	1951	1.1×10^{-3}	0.75 [0.64, 0.89]	407	6.0×10^{-8}
19p13.12	*RAB8A* ‡*	rs7252188	16228701	A (0.39)	3.2×10^{-3}	0.89 [0.82, 0.96]	1844	9.2×10^{-6}	0.69 [0.59, 0.81]	455	6.8×10^{-7}
20q12	*MAFB* ‡	rs6016392	39246610	A (0.5)	8.3×10^{-10}	1.27 [1.18, 1.37]	2004	3.9×10^{-2}	1.18 [1.01, 1.39]	452	9.9×10^{-7}
18q12.1	*MIR302F* ‡*	rs11083400	28056959	T (0.45)	CL			CP			CL & CP
					3.1×10^{-3}	0.79 [0.68, 0.92]	484	1.6×10^{-5}	1.45 [1.23, 1.72]	424	6.2×10^{-7}
10q24.33	*SH3PXD2A* ‡	rs11191818	105591779	G (0.31)	CLP			CP			CLP & CP
					2.7×10^{-4}	0.82 [0.74, 0.91]	1102	5.1×10^{-4}	1.40 [1.16, 1.69]	341	9.2×10^{-7}
1p21.3	*MIR137HG* ‡*	rs2802532	98528830	C (0.09)	CL			CLP			CL & CLP
					7.3×10^{-5}	0.59 [0.45, 0.76]	217	3.5×10^{-5}	1.41 [1.20, 1.65]	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; CI, confidence interval; 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)

https://doi.org/10.1371/journal.pgen.1009584.t001

one subtype while decreasing risk for the other (S12B and S13B Figs). The RR estimates and their 95% CIs for the top several SNPs at these loci confirmed the opposite effect of these loci on risks of OFC subtypes (Fig 2). While *SH3PXD2A* has been recently implicated in the formation of CL/P in an European GWAS [31], the *MIR302F* locus is a novel genetic risk factor for OFCs.

Locus-specific effects at regions of genetic overlap between CL/P & CP vary by racial/ethnic group

The regional association plots seem to indicate that signals of genetic overlap at the *PAX7*, *FOXE1* and *NOG* loci are driven by the European subjects; *IRF6*, *SHROOM3* and *MAFB* loci

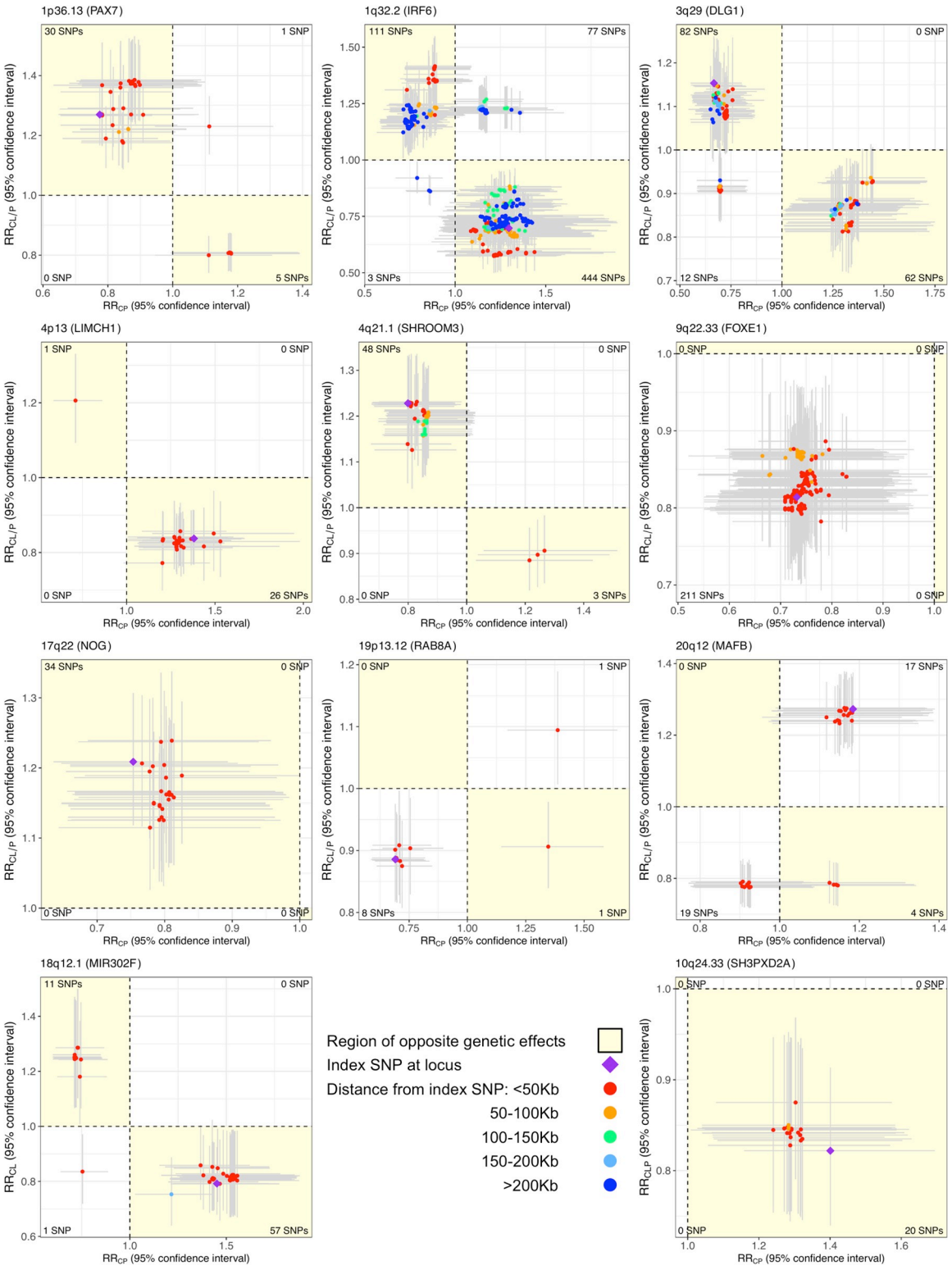

Fig 2. Scatter plot of relative risk (RR) estimates, along with corresponding 95% confidence intervals (CIs), for variants in the 9 loci showing statistical evidence of genetic overlap between CL/P & CP, along with 2 additional loci of genetic overlap between component OFC subtypes. RR estimates are 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 $> 10^{-3}$ for the purposes of this visualization. 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.

https://doi.org/10.1371/journal.pgen.1009584.g002

by the Asian subjects; while the *DLG1* and *RAB8A* loci seem to draw upon evidence from both groups (S3C–S3E, S4C–S4E, S5C–S5E, S6C–S6E, S7C–S7E, S8C–S8E, S9C–S9E, S10C–S10E and S11C–S11E Figs). Similarly, evidence for the *LIMCH1* locus seems to be driven by both Asian and Latin American subjects (Fig 3C and 3E). However, we must note that signals in

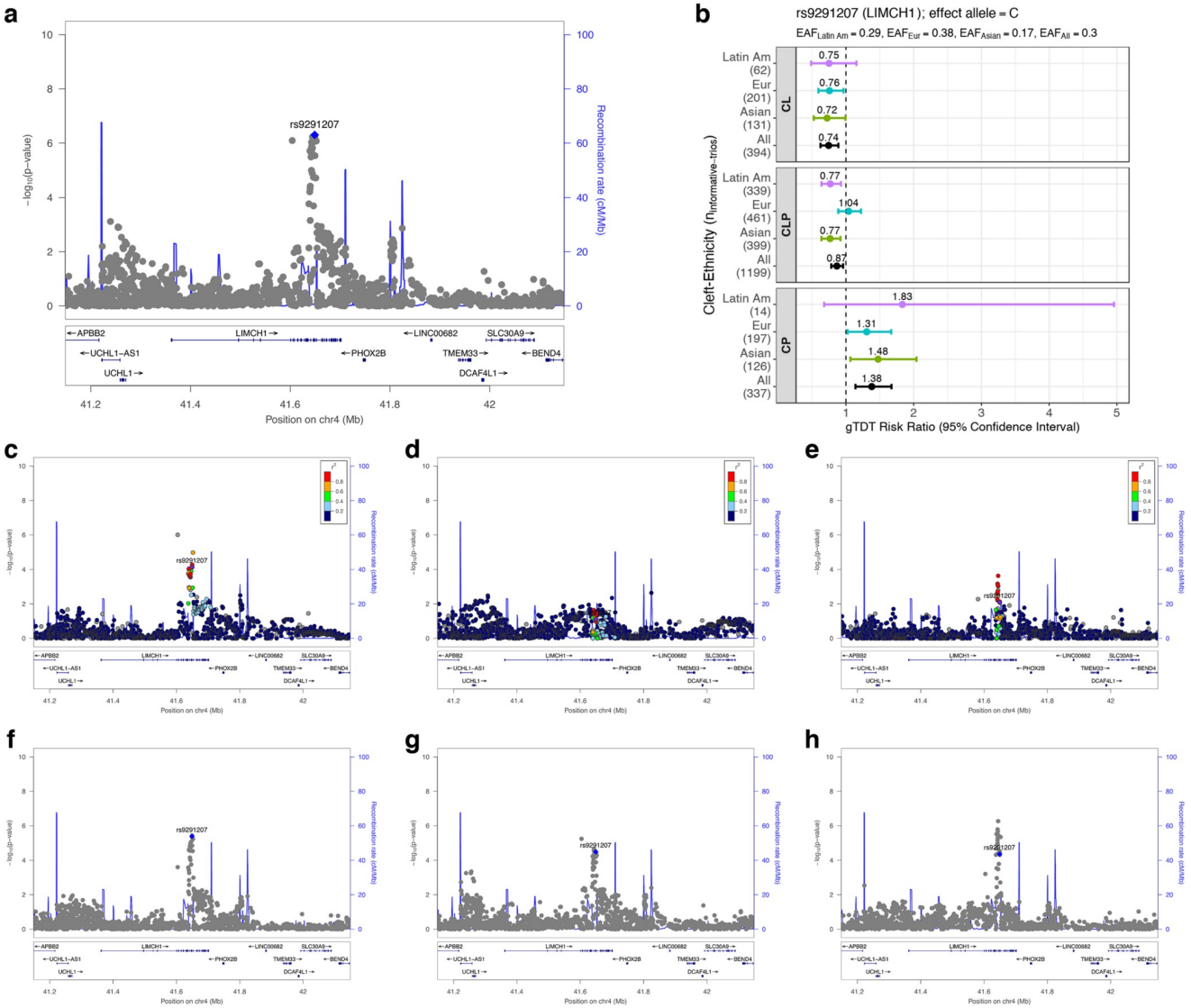

Fig 3. Regional association plots for 4p13 (*LIMCH1*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of (**a**) CL/P & CP (multi-ethnic), (**c**) CL/P & CP in Asian ancestry, (**d**) CL/P & CP in European ancestry, (**e**) CL/P & CP in Latin American 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.

https://doi.org/10.1371/journal.pgen.1009584.g003

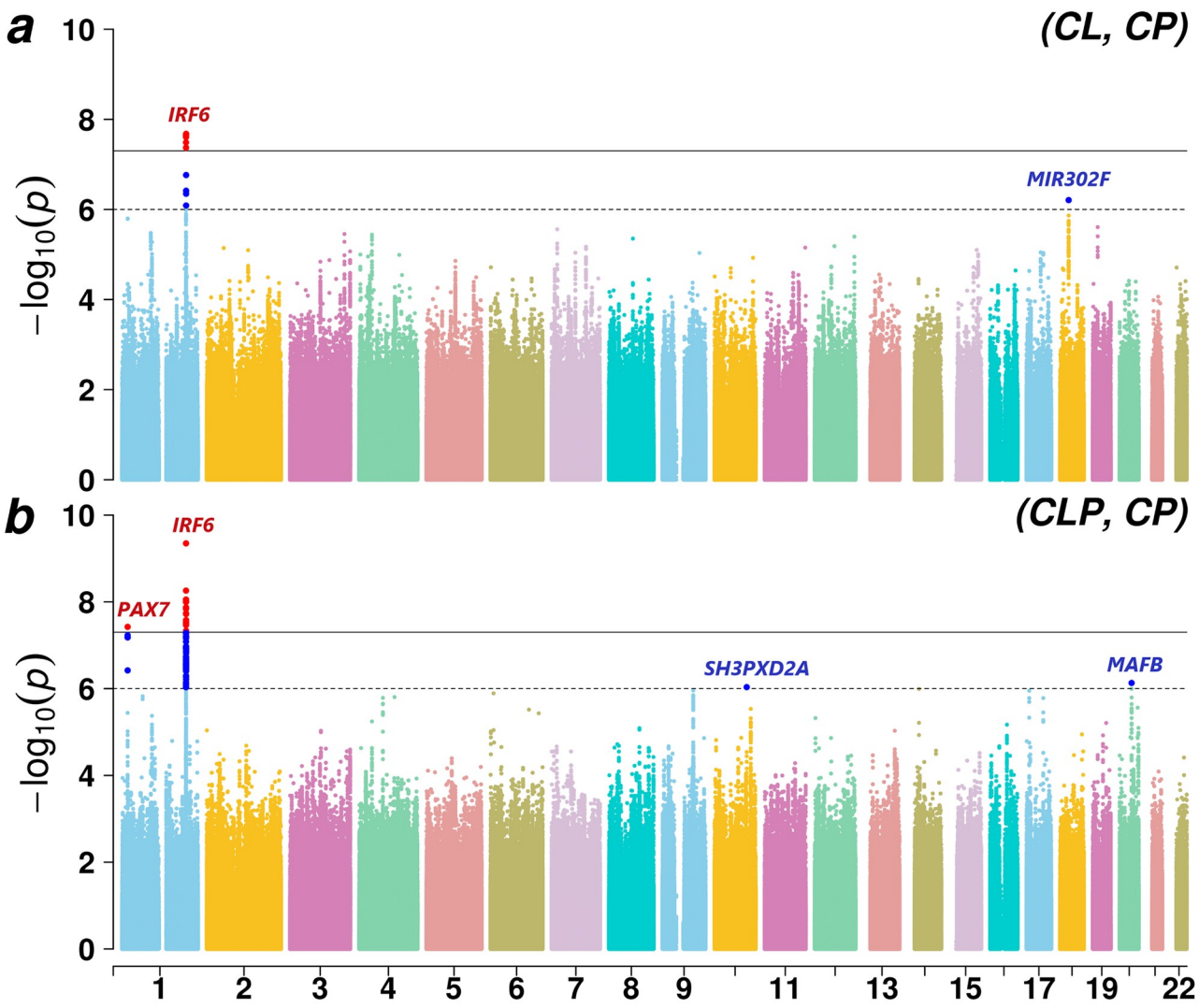

Fig 4. Manhattan plots for genome-wide analyses of genetic overlap between pairs of OFC subtypes: CL & CP, and CLP & CP. The chromosome numbers 1–22 are indicated along the x-axis. Solid black and dashed black lines are used in both plots to indicate the conventional genome-wide significance threshold 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.

https://doi.org/10.1371/journal.pgen.1009584.g004

these stratified analyses are confounded by differences in overall sample sizes between racial/ethnic groups (S1 Table), 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 (Fig 3B and S3B–S11B Figs). Notably for the Latin American subjects, the large uncertainty in the estimates for CP and the highly 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, 16]. Recent studies have indicated sex-specific differences in pleiotropic effects on complex traits [32]. 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 (S14 Fig). 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 CL/P subtypes

Analysis of subtypes CL & CLP using PLACO identified nearly all the recognized risk genes for CL/P subgroup alone as detected by the conventional gTDT analysis (Fig 5). In other words, the regions of genetic overlap identified by PLACO matched the shared signals captured by the pooled method analysis of CL and CLP subtypes. The RR estimates in S2 Table indicate the effect alleles at lead SNPs in all regions of genetic overlap affect risk of both CL and CLP in the same direction, which is consistent with the vast literature of epidemiologic and genome-wide studies of OFCs [6–8, 12, 14, 21, 30].

Beyond shared etiology, few loci of genetic overlap exhibit etiologic differences between CL and CLP

When investigating the RR estimates from the top several SNPs at each of the above-mentioned loci shared by CL and CLP, we found loci suggesting different pathogenesis of these subtypes at 1p22.1 (*ABCA4*, *ARHGAP29*) and 1q32.2 (*IRF6*) as evidenced by a large number of SNPs with opposite genetic effects (S15 Fig). Some previous studies [14, 33] have noted differences in genetic etiologies of CL and CLP, particularly at the *IRF6* locus. Additionally, the 1p36.13 (*PAX7*), 3q12.1 (*COL8A1*), 8q21.3 (*DCAF4L2*) and 8q24 (gene desert) regions with shared effects between CL and CLP appear to have at least 2 distinct loci of genetic overlap

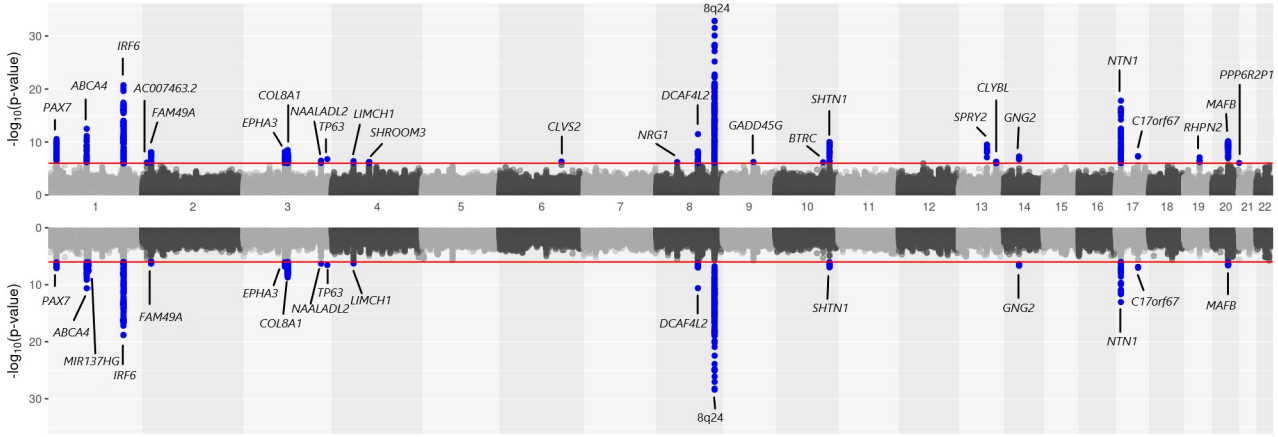

Fig 5. Mirrored Manhattan plot of genome-wide analyses of CL/P, and CL & CLP. Results from genetic associations of CL/P using the gTDT are shown in the upper panel. Results from shared genetic associations between CL & CLP using PLACO are shown in the lower panel. The red horizontal lines in the two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

https://doi.org/10.1371/journal.pgen.1009584.g005

indicated by more than one peaks (S16 Fig). Presence of possibly independent loci in the 8q24 region has been reported previously [14]. Furthermore, PLACO revealed a novel OFC locus at 1p21.3 (*MIR137HG*, $p = 2.2 \times 10^{-8}$) with opposite genetic effects for CL & CLP at the genome-wide threshold (Table 1 and S17A Fig). The estimated RRs of the effect allele at the lead SNP of this locus indicate its protective effect on CL, and its deleterious effect on subtypes CLP and CP across racial/ethnic groups (S17B Fig). This signal is, however, lost from CL/P (S17C Fig) due to pooling together of opposite effects of variants on CL and CLP (S17D and S17E Fig). Consequently, this locus near *MIR137HG* fails to show any evidence of genetic overlap between CL/P & CP (S17F Fig) even though there is moderately strong statistical evidence of genetic overlap between all pairs of OFC subtypes (S17A, S17G and S17H Fig).

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 here the intuition and statistical model behind PLACO. For a variant with relative risks RR_1 and RR_2 on two disease subgroups, three possible situations can arise: *global null*, where the variant has no genetic effect on either subgroup ($RR_1 = 1$, $RR_2 = 1$); *sub-null*, where the variant influences risk to one subgroup but not the other (either $RR_1 = 1$, $RR_2 \neq 1$ or $RR_1 \neq 1$, $RR_2 = 1$); and finally *non-null*, where the variant influences risk to both subgroups ($RR_1 \neq 1$, $RR_2 \neq 1$). 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 Material and 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 situation exist. Note, meta-analysis techniques such as inverse-variance weighted meta-analysis or Fisher's combination of p-values [34] may be employed here, but just like the pooled method they test the global null hypothesis, which is not exactly the null hypothesis to test when the goal is to identify common genetic basis of two disease subgroups.

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 (S18A Fig). 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 (S18B and S18C Fig). 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 (S19A–S19C Fig). This observation holds true irrespective of how widely different the MAFs are between the two simulated ethnic groups (S19D–S19F Fig). 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. When we included other potential methods (e.g., different meta-analysis techniques) for comparison with PLACO, we found the meta-analysis methods are comparable to the pooled method and show similar lack of type I error control (S1 Appendix). 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 naïve 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 naïve eapproaches: 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 naïve approaches (S20 Fig). 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% SNP influences risk to both CL/P and CP with RR = 1.5. When the effects are in opposite directions ($RR_{CL/P} = 1.5$, $RR_{CP} = 1/1.5$), we found subgroup-specific p-values were $p_{CL/P} < 5 \times 10^{-8}$ and $p_{CP} < 10^{-3}$ in only 14% of our simulated data while PLACO satisfied $p_{PLACO} < 5 \times 10^{-8}$ in 44% of the same data. This indicates the improved power of PLACO over relying on subgroup-specific p-values in identifying genetic overlap. 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 (S20 Fig).

Discussion

In this analysis of multi-ethnic case-parent trios from the POFC and the GENEVA studies, we identified genetic overlap between nonsyndromic CL/P & CP at 1 locus in 1q32.2 reaching genome-wide significance (5×10^{-8}), 2 loci in 1p36.13 and 17q22 barely missing this conventional threshold, and 6 loci in 3q29, 4p13, 4q21.1, 9q22.33, 19p13.12 and 20q12 yielding suggestive significance at a threshold of 10^{-6}. The apparent risk SNPs at 4p13 and 19p13.12 are located in the *LIMCH1* and *RAB8A* genes respectively, which have not been implicated in OFC genetics before. We found evidence of shared etiology at 3 of these 9 loci, including the well-recognized *FOXE1* gene which influences risk to both CL/P and CP in the same direction. The effect alleles at the other loci found in this analysis appear to increase risk for one OFC subgroup but decrease risk for the other. We do not discuss the GWAS findings for each cleft subgroup separately since they have been described elsewhere [12, 14]. At the suggestive significance level, we further identified genetic overlap between CL & CP, CLP & CP, and CL & CLP at 3 loci in 18q12.1, 10q24.3 and 1p21.3 respectively, of which the loci 18q12.1 near *MIR302F* and 1p21.3 near *MIR137HG* have not been previously shown to be associated with risk to OFCs. We replicated shared etiology of CL & CLP in/near several recognized genes, and further found opposite effects of top several SNPs at a few loci hinting at potentially different pathogenesis of these 2 OFC subtypes. None of the loci identified in this study appear to exhibit sex-dependent genetic overlap.

While *LIMCH1*, *RAB8A*, *MIR302F* and *MIR137HG* are novel risk genes for OFCs, all the other loci we identified are in/near genes previously implicated in GWAS of CL/P [2, 14]. For instance, *PAX7* has been found in multiple GWAS of CL/P including a coding *de novo* variant [35]. *IRF6*, *FOXE1* and *NOG* have also been found in multiple GWAS, with additional

evidence of functional common variants found in experimentally validated enhancer regions in these loci [12, 35–37]. Association of *DLG1* with CL/P has been reported in a recent GWAS [38] and experimentally validated in a mouse model [39]. Similarly, mouse experiments found cleft and neural tube defects (NTDs) following alterations of the *SHROOM3* gene [39], and *MAFB* has been identified in GWAS of CLP with mouse models showing its role in palate development [16].

The associated SNPs at the 4p13 locus occur within a topologically associating domain containing two genes: *LIMCH1* and *UCHL1* [40]. This locus was previously found to be suggestively associated with CL/P using GENEVA and POFC datasets [14]. 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 NTDs [41]. Although OFCs and NTDs are considered 'mid-line birth defects', and supplementing mothers with folate and multivitamins during pregnancy seem to reduce risk to both [42], 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 transcription [43]. 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 Belgium [31]. Zebrafish and mouse models support some role of this gene in OFCs [31]. Homozygous disruption of this gene in mutant mice resulted in complete clefts of the secondary palate [44]. 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 syndrome [16]. 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 models [45]. 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 development as the other members of the *MIR302* family.

In this manuscript we have annotated the 1p21.3 and 19p13.12 loci with the genes *MIR137HG* and *RAB8A*, respectively. However, these annotations are based on proximity to the most significant SNPs and there is no specific evidence in the literature to support their role in craniofacial development. There are 3 genes in the topological domain containing the associated SNPs of 1p21.3 locus (*MIR137HG*, *MIR2682* and *DPYD*) [40], which may be associated with schizophrenia and bipolar disorder [46]. *RAB8A* itself has been shown to be associated with endometrial cancer [47]. It will be an area of future work to replicate and further elucidate our findings near *RAB8A* and *MIR137HG*.

Taken together, our study provides strong statistical evidence for possible overlap in the genetic architectures of CL/P and CP. We have leveraged multi-ethnic case-parent trios with children affected by a nonsyndromic OFC. One of the great advantages of family studies in general, and case-parent trio designs in particular, is that inference based on deviation from independent assortment of parental alleles is not subject to population stratification, and therefore type I error inflation is not a concern in the TDT-based inference presented here. Biases can still arise in trio-based studies due to genotyping or imputation errors, leading to

Mendelian inconsistencies and subsequent bias [48]. For this particular reason, we took family relationship into account when phasing the genetic data before imputation, which was shown to greatly alleviate these concerns [48]. We examined the QQ plots (S1 and S2 Figs), which further corroborated that wide-spread type I error inflation does not exist. There might be concerns about spurious findings for the loci with opposite genetic effects. We note that, under the null, getting a spurious signal at a locus with opposite effects is equally likely as effects in the same direction. We allay concerns about spurious results from low allele frequencies by focusing only on common variants. Further, we plotted confidence intervals of top several SNPs at each discovered locus (Fig 2 and S15 Fig), which alleviate concerns about large uncertainties in estimated effect sizes and directions.

Historically, CL/P and CP have been thought to have distinct etiologies based on developmental origins and epidemiologic patterns. Linkage studies first identified significant evidence of linkage for markers in the *FOXE1* region in CL/P multiplex families [49]; subsequent studies confirmed this gene as a risk factor for both CL/P and CP [11, 50]. Linkage analysis identified a susceptibility locus near *TBX22* for CL/P and a later study found mutations in *TBX22* in CP individuals [18]. Fine-mapping of translocation breakpoints revealed an important role of *SATB2* in cases with CP; a few years later, a candidate gene study identified significant association of a variant in *SATB2* with CL/P in two Asian populations [10, 18]. There exists some evidence that variants in *IRF6* [29], *GRHL3* [29], *ARHGAP29* [51–53] and *MSX1* [18] regions may affect risk to both CL/P and CP in the same direction (often termed as 'shared genetic risk'). Note, much of this evidence are from patterns of Mendelian inheritance of rare variants in extended pedigrees; genetic overlap of common variants in nonsyndromic OFCs may not follow the same patterns. In recent GWAS, variants near *IRF6* [21] and *NOG* [14, 20] have shown weak evidence of decreased risk for one OFC subgroup and increased risk for the other. Among these well-recognized risk genes for OFCs, only *FOXE1* has been successfully implicated as a shared risk gene in GWAS [12]. Our method PLACO not only replicated this finding for *FOXE1* at a suggestive threshold of 10^{-6}, 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* (S21 Fig). It is possible that rarer variants drive genetic overlap in these regions, and PLACO is currently only applicable to common variants. Since we are interested in identifying shared genetic risk at the variant level, we have not considered any set- or gene-based testing. Another limitation is that we have only assessed genetic sharing using estimates of direct effects of inherited genotypes of offspring on their disease status, and did not consider any indirect effect (e.g., maternal effect). We did not consider SNP × SNP interactions due to substantially reduced power of such tests. We obtained variant-level summary statistics using the gTDT, which cannot handle incomplete trios; however, if there is a large number of dyads along with trios, one may obtain the summary statistics using the generalized disequilibrium test [54].

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.

Material and methods

Ethics statement

The GENEVA study 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 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 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, 51]. Type of cleft, sex, race as well as common environmental risk factors were obtained through direct maternal interview [51].

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, 55]. 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 distribution of trios by cleft subtype and racial/ethnic groups from both studies is given in S1 Table. It is to be noted that originally 412 individuals from POFC were included in GENEVA [55]; 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 approach [56].

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 Server [57] 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 SHAPEIT2 [58], and original genotyped SNPs on build hg18 were lifted over to hg19 using 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 results [48]. 'Hard' genotype calls were made by setting threshold 0.1 within PLINK 2.0 [59]. 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. All variants are in the forward strand. We took the following quality control measures: all genotyped SNPs with missingness >5% and Mendelian error rate >5% were removed. All SNPs with MAF< 1% and 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 [60]. Additionally, individuals with low genotype information or evidence of low-quality DNA, individuals with SNP missingness >10%, individuals duplicated across the POFC and GENEVA datasets, and individuals with incomplete trio status 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 ≥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 IMPUTE2 [61] to the 1000 Genomes Phase 3 reference panel as described previously [55]. 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 ≥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.

Overview of 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 [62, 63] 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 [24, 25] 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 associated [28]. 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 previously [28]. The PLACO test statistic is

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

and its approximate, asymptotic p-value is given by

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

where z_1 and z_2 are the observed Z-scores for the two disorders at any given SNP; $\mathbb{F}(u) = 2\int_{|u|}^{\infty} \mathbb{f}(x)dx$ 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 model [28]. Open-source implementation of PLACO in R [64] is available in GitHub (see Web Resources). While PLACO was originally proposed for two traits from population-based studies (e.g. case-control traits) [28], here we showed PLACO can very well be used for family studies as long as the summary statistics are obtained after appropriately accounting for all confounding effects, including relatedness and population stratification. It is worth noting that the traditional GWAS thresholds are still appropriate for PLACO although the null hypothesis of PLACO and its interpretation is different from a traditional GWAS null hypothesis. This is because PLACO is a bivariate statistical method that does not do any multiple comparison across traits. As a result, the number of multiple comparisons one is doing when applying PLACO genome-wide on two traits is the same as the number of multiple comparisons one does when performing a traditional GWAS on a single trait.

Statistical analyses

For all analyses presented here, we focused on bi-allelic SNPs with MAF≥5%, where MAF is calculated based on only the parents (founders) using PLINK 2.0. For each study separately, we obtained summary statistics of genetic association between each variant and each OFC subgroup using the gTDT model for case-parent trios under an additive genetic model as implemented in R package `trio` [62] (v3.20.0). We used the gTDT over the allelic TDT because of its several advantages [65]: gTDT can be more powerful; yields parameter estimates, standard errors along with p-values; and enables direct assessment of RR. However, unlike the allelic TDT [66], the gTDT assumes a specific mode of inheritance; here we chose an additive model. The gTDT in `trio` package uses the minor allele of the input dataset as the effect allele. Since

the gTDT is applied on each subgroup and each study separately, it is possible that a minor allele is not the same across subgroups and across studies. Therefore, we set the minor allele from the CL/P subgroup in POFC as the effect allele for all analyses. We, then, meta-analyzed the gTDT results over the two studies using inverse-variance weighted fixed effects meta-analysis. We implemented PLACO v0.1.1 on the meta-analyzed gTDT results from subgroups CL/P and CP to identify possible genetic overlap between them. We used the default parameter choices as described in S1 Appendix. To identify regions of significant genetic overlap, we used the conventional genome-wide threshold 5×10^{-8} and also a suggestive threshold of 10^{-6}. In this paper, we do not discuss the GWAS findings for CL/P or CP separately; they have been described in prior manuscripts using previously-imputed GENEVA and POFC data [12, 14].

We explored if any identified region of genetic overlap was modified by sex. To do this, we first obtained summary statistics from the 1 df SNP×Sex analysis using the gene-environment gTDT model in `trio` package (again assuming additive genetic model); then meta-analyzed the SNP×Sex estimates across the two studies; and finally applied PLACO on the meta-analyzed 1 df SNP×Sex summary statistics. For each analysis, we created Manhattan plots to show signals, and QQ plots to check for potential bias in the association results. For the QQ plots, we also calculated the genomic inflation factors at the 50^{th} percentile ($\lambda_{0.5}$) to quantify the extent of the bulk inflation, and at the $1/10^{th}$ of a percentile ($\lambda_{0.001}$) to quantify inflation towards the meaningful tail of the distribution. We took the p-values from a given method (e.g., gTDT, PLACO), mapped them to 1 df χ^2 statistics, and calculated λ_x as the ratio of empirical $100(1-x)^{th}$ percentile of these statistics and the theoretical $100(1-x)^{th}$ percentile of 1 df χ^2 distribution.

Stratified analyses

We considered two stratified analyses: one stratified by racial/ethnic group and the other by cleft subtype (CL, CLP, CP). As described before, genome-wide meta-analyzed gTDT summary statistics were obtained for CL/P and for CP within each of the three major racial/ethnic groups: European, Asian and Latin American. PLACO was applied on CL/P and CP summary data within each racial/ethnic group. For the OFC subtype stratified analyses, meta-analyzed gTDT summary statistics were obtained for each cleft subtype, and then PLACO was applied on each of the three pairwise combinations to compare and contrast results against those from the main analysis. Note, we estimated power of such an analysis by performing some simulation experiments mimicking the lead SNPs of two loci of genetic overlap (see S1 Appendix).

Locus annotation and candidate gene prioritization

For each analysis of genetic overlap, we defined independent loci by clumping all the suggestively significant SNPs ($p_{\text{PLACO}} < 10^{-6}$) in a ±500 Kb radius and with linkage disequilibrium (LD) $r^2 > 0.2$ into a single genetic locus. This clumping was done using `FUMA` [67] (`SNP2GENE` function, v1.3.5e). Since we performed multi-ethnic analysis, we separately used 1000G Phase 3 EUR, EAS and AMR as reference populations for LD calculation. The number of independent loci and the index SNP for each locus (chosen to be the most significant SNP) were the same regardless of the racial/ethnic groups assumed for LD calculation. To define the bounds of each locus, as used in the regional association plots annotated by effect size directions, 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 LocusZoom [68] 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 datasets; combining samples improves power to detect genetic associations over a two-stage discovery- replication approach [56]. 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 on OFC trios; and (3) an assessment of our findings based on existing literature and large publicly available databases. 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.

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 naïve 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. Refer S1 Appendix for more details on our simulation experiments.

Genome build

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

Supporting information

S1 Appendix. Additional text and supporting information. It includes additional details on simulation experiments, power calculations, and PLACO implementation.
(PDF)

S1 Fig. QQ plots of the gTDT summary statistics used for different genetic overlap analysis using PLACO in this manuscript.
(PDF)

S2 Fig. QQ plots from the different PLACO analyses, including stratified analyses, conducted in this manuscript.
(PDF)

S3 Fig. Regional association plots for 1q32.2 (*IRF6*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S4 Fig. Regional association plots for 3q29 (*DLG1*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S5 Fig. Regional association plots for 17q22 (*NOG*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S6 Fig. Regional association plots of PLACO p-values, annotated by directions of effect sizes, 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. Index SNP here is the lead (most significant) SNP at each locus. SNPs with opposite genetic effects for 2 OFC subgroups are colored in golden yellow while those with shared effects are in dark green. The directions of genetic effects are determined from the relative risk (RR) estimates for each subgroup as provided by the gTDT method. RR estimates and the corresponding 95% confidence intervals for the SNPs above the dashed red horizontal line are portrayed in Fig 2.
(PDF)

S7 Fig. Regional association plots for 1p36.13 (*PAX7*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S8 Fig. Regional association plots for 4q21.1 (*SHROOM3*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S9 Fig. Regional association plots for 9q22.33 (*FOXE1*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S10 Fig. Regional association plots for 20q12 (*MAFB*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S11 Fig. Regional association plots for 19p13.12 (*RAB8A*) identified as a region of genetic overlap between CL/P & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(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, **(G)** CLP & CP, **(H)** CL & CLP. 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 most strongly associated SNP, 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 and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S12 Fig. Regional association plots for 18q12.1 (*MIR302F*) identified as a region of genetic overlap between CL & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(C)** CL & CP, **(D)** CLP & CP, **(E)** CL & CLP. The blue diamond represents the most strongly associated SNP in the region showing evidence of genetic overlap. For multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used to represent strength of LD. Panel **(B)** shows relative risk estimates and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S13 Fig. Regional association plots for 10q24.33 (*SH3PXD2A*) identified as a region of genetic overlap between CLP & CP. LocusZoom plots focus on PLACO analysis of **(A)** CL/P & CP, **(C)** CL & CP, **(D)** CLP & CP, **(E)** CL & CLP. The blue diamond represents the most strongly associated SNP in the region showing evidence of genetic overlap. For multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used to represent strength of LD. Panel **(B)** shows relative risk estimates and their 95% confidence intervals as obtained from the gTDT analyses.
(PDF)

S14 Fig. Regional association plots to investigate if genetic overlap between CL/P & CP at any region of interest (as identified by PLACO in different pairwise analyses) is modified by sex. P-values from our SNP × Sex analyses are plotted. The blue diamond represents the most strongly associated SNP in the region of genetic overlap. For multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used to represent strength of LD.
(PDF)

S15 Fig. Scatter plot of relative risk (RR) estimates of CL and CLP, along with corresponding 95% confidence intervals (CIs), for variants in the 26 loci for CL/P. 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. While the 26 loci were identified from the gTDT analysis of CL/P at a suggestive threshold of 10^{-6}, the SNPs plotted here are selected based on two criteria: (i) SNPs in ±500 Kb radius and in LD $r^2 > 0.2$ with the index SNP (the most significant SNP in the locus); and (ii) SNPs with PLACO p-value $<10^{-3}$ from the genetic overlap analysis of CL & CLP. Two loci, 6q22.31 and 19q13.11 (see S2 Table), are not depicted here since no SNP in these 2 loci show significance in PLACO analysis even at a liberal threshold of 10^{-3}. These plots show the concordance of results for CL/P, and CL & CLP; i.e., the regions of genetic overlap identified by PLACO matches with the shared signals captured by the pooled analysis of CL and CLP subtypes. Additionally, there is some indication of genetically distinct etiology of subtypes CL and CLP

at 1p22.1 (*ABCA4*, *ARHGAP29*) and 1q32.2 (*IRF6*), as depicted by the SNPs with opposite genetic effects.
(PDF)

S16 Fig. Regional association plots of PLACO p-values from CL & CLP analysis, annotated by directions of effect sizes, for variants in some of the loci from the 26 loci for CL/P. Index SNP here is the lead (most significant) SNP from the gTDT analysis of CL/P. SNPs with opposite genetic effects for the 2 OFC subtypes are colored in golden yellow while those with shared effects are in dark green. The directions of genetic effects are determined from the relative risk (RR) estimates for each subtype as provided by the gTDT method. RR estimates and the corresponding 95% confidence intervals for the SNPs above the dashed red horizontal line are portrayed in S15 Fig. Each of the loci here appear to have at least 2 distinct regions of genetic overlap.
(PDF)

S17 Fig. Regional association plots for 1p21.3 (*MIR137HG*) identified as a region of genetic overlap between CL & CLP. LocusZoom plots focus on PLACO analysis of **(A)** CL & CLP, **(E)** CL & CLP annoted by directions of genetic effects, **(F)** CL/P & CP, **(G)** CL & CP, **(H)** CLP & CP, while **(C)** shows pooled method analysis of CL and CLP (i.e., gTDT analysis of subgroup CL/P). The blue diamond represents the most strongly associated SNP in the region showing evidence of genetic overlap. For multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used to represent strength of LD. Panel **(B)** shows relative risk (RR) estimates and their 95% confidence intervals (CIs) across OFC subtypes and racial/ethnic groups as obtained from the gTDT analyses. Panel **(D)** shows scatter plot of RR estimates of CL and CLP, along with corresponding 95% CIs, for the top several SNPs. SNPs represented here are in ±500 Kb radius and in LD $r^2 > 0.2$ with the index SNP (the most significant SNP in the locus), and with PLACO p-value $< 10^{-3}$ from the genetic overlap analysis of CL & CLP. 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. Horizontal (vertical) error bar around each RR estimate corresponds to the 95% CI for CLP (CL). The region depicting opposite genetic effects of SNPs for the 2 OFC subtypes is shaded in yellow. The number of SNPs in each quadrant is printed in the corresponding corner of the plot.
(PDF)

S18 Fig. Scenario I: QQ plots for null data from two independent bi-ethnic case-parent trio studies of OFC subgroups assuming fixed genetic effects. Observed($-\log_{10}$p-values) are plotted on the y-axis and Expected($-\log_{10}$p-values) on the x-axis. Type I error performance of tests of simultaneous effect of a genetic variant on both outcomes is based on 9.99 million null variants with genetic effects that are either $\{RR_{CL/P} = 1, RR_{CP} = 1\}$ or $\{RR_{CL/P} = 1, RR_{CP} = 1.15\}$ or $\{RR_{CL/P} = 1.15, RR_{CP} = 1\}$. The gray shaded region represents a conservative 95% confidence interval for the expected distribution of p-values. P-values $\geq 10^{-12}$ are shown here.
(PDF)

S19 Fig. Scenario II: QQ plots for null data from two independent bi-ethnic case-parent trio studies of OFC subgroups assuming fixed genetic effects for one trait and random for the other. Observed($-\log_{10}$p-values) are plotted on the y-axis and Expected($-\log_{10}$p-values) on the x-axis. Type I error performance of tests of simultaneous effect of a genetic variant on both outcomes is based on 9.99 million null variants with genetic effects $\{\log(RR_{CL/P}) \sim N(0, 0.1^2), RR_{CP} = 1\}$. Values of $RR_{CL/P}$ ranged from $\frac{1}{1.7}$ to 1.7. The gray shaded region represents a

conservative 95% confidence interval for the expected distribution of p-values. P-values $\geq 10^{-12}$ are shown here.
(PDF)

S20 Fig. Power of PLACO, pooled method, and naive approaches at genome-wide significance level (5×10^{-8}) for varying genetic effects of the two independent bi-ethnic case-parent trio studies of OFC subgroups. The first naive approach ('Naive-1') declares pleiotropic association when $p_{CL/P} < 5 \times 10^{-8}$ and $p_{CP} < 10^{-5}$, while the second naive approach ('Naive-2') uses a more liberal criterion $p_{CL/P} < 5 \times 10^{-8}$ and $p_{CP} < 10^{-3}$. Note, unlike PLACO, the pooled method lacks type I error control in most scenarios of sample size and/or MAF imbalance, and hence its power should be interpreted with caution.
(PDF)

S21 Fig. Regional association plots to investigate if genetic overlap between CL/P & CP is identified by PLACO at some candidate regions based on past literature. The blue diamond represents the most strongly associated SNP in the region of genetic overlap. For multi-ethnic analyses, there is no unique LD between SNPs and hence no color has been used to represent strength of LD.
(PDF)

S1 Table. Distribution of independent complete case-parent trios in the POFC and the GENEVA studies by racial/ethnic group and by cleft subtypes.
(PDF)

S2 Table. Association results for the most significant markers from the 26 loci for CL/P at a suggestive threshold of 10^{-6}. The analysis of CL/P is the same as pooled analysis of CL and CLP subtypes. Results from the genetic overlap analysis of CL & CLP using PLACO at these loci are also provided. Analyses are based on all trios from both POFC and GENEVA for a given cleft subtype. The "No. of trios" columns give the numbers of complete informative case-parent trios as used by gTDT. Note, the relevant 95% confidence intervals for the gTDT relative risk (RR) estimates here are provided in S15 Fig. CL & CLP PLACO p-value $< 5 \times 10^{-8}$ for a SNP indicates statistically significant association of the SNP with both CL and CLP at the genome-wide level. PLACO p-value $< 10^{-6}$ is considered suggestive evidence of genetic overlap.
(PDF)

Acknowledgments

This research was carried out using computing cluster—the Joint High Performance Computing Exchange—at the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health.

Web Resources

BCFtools, https://samtools.github.io/bcftools
FUMA, https://fuma.ctglab.nl/
liftOverPlink, http://github.com/sritchie73/liftOverPlink
Locuszoom plot, http://locuszoom.org/
Manhattan plot, https://github.com/YinLiLin/CMplot
PLACO, https://github.com/RayDebashree/PLACO/
PLINK, https://www.cog-genomics.org/plink/2.0
trio, https://bioconductor.org/packages/release/bioc/html/trio.html
3D Genome Browser, http://promoter.bx.psu.edu/hi-c/

Author Contributions

Conceptualization: Debashree Ray, Terri H. Beaty.

Data curation: Sowmya Venkataraghavan, Wanying Zhang, Jacqueline B. Hetmanski.

Formal analysis: Debashree Ray, Sowmya Venkataraghavan, Wanying Zhang.

Funding acquisition: Debashree Ray.

Investigation: Debashree Ray, Elizabeth J. Leslie, Ingo Ruczinski, Margaret A. Taub, Terri H. Beaty.

Methodology: Debashree Ray, Ingo Ruczinski.

Project administration: Debashree Ray, Terri H. Beaty.

Resources: Seth M. Weinberg, Jeffrey C. Murray, Mary L. Marazita, Terri H. Beaty.

Software: Debashree Ray, Sowmya Venkataraghavan, Wanying Zhang.

Supervision: Debashree Ray, Terri H. Beaty.

Validation: Debashree Ray, Sowmya Venkataraghavan, Elizabeth J. Leslie, Margaret A. Taub.

Visualization: Debashree Ray, Sowmya Venkataraghavan.

Writing – original draft: Debashree Ray.

Writing – review & editing: Debashree Ray, Sowmya Venkataraghavan, Wanying Zhang, Elizabeth J. Leslie, Jacqueline B. Hetmanski, Seth M. Weinberg, Mary L. Marazita, Ingo Ruczinski, Margaret A. Taub, Terri H. Beaty.

References

1. Wehby G, Cassell CH. The impact of orofacial clefts on quality of life and healthcare use and costs. Oral Dis. 2010; 16(1):3–10. https://doi.org/10.1111/j.1601-0825.2009.01588.x PMID: 19656316
2. Beaty TH, Marazita ML, Leslie EJ. Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. F1000Research. 2016; 5. https://doi.org/10.12688/f1000research.9503.1 PMID: 27990279
3. Christensen K, Mortensen PB. Facial clefting and psychiatric diseases: a follow-up of the Danish 1936–1987 Facial Cleft cohort. Cleft Palate Craniofac J. 2002; 39(4):392–396. https://doi.org/10.1597/1545-1569_2002_039_0392_fcapda_2.0.co_2 PMID: 12071787
4. Leslie EJ, Marazita ML; Wiley Online Library. Genetics of cleft lip and cleft palate. Am J Med Genet Part C Semin Med Genet. 2013; 163C(4):246–258. https://doi.org/10.1002/ajmg.c.31381 PMID: 24124047
5. Christensen K, Juel K, Herskind AM, Murray JC. Long term follow up study of survival associated with cleft lip and palate at birth. BMJ. 2004; 328:1405. https://doi.org/10.1136/bmj.38106.559120.7C PMID: 15145797
6. Sperber GH. Palatogenesis: Closure of the Secondary Palate. In: Wyszynski DF, editor. Cleft lip and palate: from origin to treatment. Oxford: Oxford University Press; 2002. p. 5–24.
7. Sivertsen Å, Wilcox AJ, Skjærven R, Vindenes HA, Åbyholm F, Harville E, et al. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. BMJ. 2008; 336:432–434. https://doi.org/10.1136/bmj.39458.563611.AE PMID: 18250102
8. Mangold E, Ludwig KU, Nöthen MM. Breakthroughs in the genetics of orofacial clefting. Trends Mol Med. 2011; 17(12):725–733. https://doi.org/10.1016/j.molmed.2011.07.007 PMID: 21885341
9. Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, et al. Mutations in *IRF6* cause Van der Woude and popliteal pterygium syndromes. Nat Genet. 2002; 32(2):285–289. https://doi.org/10.1038/ng985 PMID: 12219090
10. Beaty T, Hetmanski J, Fallin M, Park J, Sull J, McIntosh I, et al. Analysis of candidate genes on chromosome 2 in oral cleft case-parent trios from three populations. Hum Genet. 2006; 120(4):501–518. https://doi.org/10.1007/s00439-006-0235-9 PMID: 16953426

11. Moreno LM, Mansilla MA, Bullard SA, Cooper ME, Busch TD, Machida J, et al. *FOXE1* association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. Hum Mol Genet. 2009; 18 (24):4879–4896. https://doi.org/10.1093/hmg/ddp444 PMID: 19779022

12. Leslie EJ, Carlson JC, Shaffer JR, Butali A, Buxó CJ, Castilla EE, 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. 2017; 136(3):275–286. https://doi.org/10.1007/s00439-016-1754-7 PMID: 28054174

13. Ludwig KU, Böhmer AC, Bowes J, Nikolić M, Ishorst N, Wyatt N, et al. Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. Hum Mol Genet. 2017; 26(4):829–842. https://doi.org/10.1093/hmg/ddx012 PMID: 28087736

14. Carlson JC, Anand D, Butali A, Buxo CJ, Christensen K, Deleyiannis F, et al. A systematic genetic analysis and visualization of phenotypic heterogeneity among orofacial cleft GWAS signals. Genet Epidemiol. 2019; 43:704–716. https://doi.org/10.1002/gepi.22214 PMID: 31172578

15. He M, Zuo X, Liu H, Wang W, Zhang Y, Fu Y, et al. Genome-wide Analyses Identify a Novel Risk Locus for Nonsyndromic Cleft Palate. J Dent Res. 2020; 10:1–8. PMID: 32758111

16. Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. Nat Rev Genet. 2011; 12(3):167–178. https://doi.org/10.1038/nrg2933 PMID: 21331089

17. Marazita M, Leslie E. Genetics of nonsyndromic clefting. In: Losee JE, Kirschner RE, editors. Comprehensive Cleft Care. Boca Raton FL: CRC Press; 2016. p. 207–224.

18. Rahimov F, Jugessur A, Murray JC. Genetics of nonsyndromic orofacial clefts. Cleft Palate Craniofac J. 2012; 49(1):73–91. https://doi.org/10.1597/10-178 PMID: 21545302

19. Carlson JC, Taub MA, Feingold E, Beaty TH, Murray JC, Marazita ML, et al. Identifying genetic sources of phenotypic heterogeneity in orofacial clefts by targeted sequencing. Birth Defects Res. 2017; 109 (13):1030–1038. https://doi.org/10.1002/bdr2.23605 PMID: 28762674

20. Moreno Uribe L, Fomina T, Munger R, Romitti P, Jenkins M, Gjessing HK, et al. A population-based study of effects of genetic loci on orofacial clefts. J Dent Res. 2017; 96(11):1322–1329. https://doi.org/10.1177/0022034517716914 PMID: 28662356

21. Huang L, Jia Z, Shi Y, Du Q, Shi J, Wang Z, et al. Genetic factors define CPO and CLO subtypes of nonsyndromic orofacial cleft. PLoS Genet. 2019; 15(10):e1008357. https://doi.org/10.1371/journal.pgen.1008357 PMID: 31609978

22. Howe LJ, Lee MK, Sharp GC, Smith GD, St Pourcain B, Shaffer JR, et al. Investigating the shared genetics of non-syndromic cleft lip/palate and facial morphology. PLoS Genet. 2018; 14(8):e1007501. https://doi.org/10.1371/journal.pgen.1007501 PMID: 30067744

23. Carlson JC. Methods for family-based designs in genetic epidemiology studies. University of Pittsburgh; 2017.

24. Ray D, Chatterjee N. Effect of non-normality and low count variants on cross-phenotype association tests in GWAS. Eur J Hum Genet. 2020; 28(3):300–312. https://doi.org/10.1038/s41431-019-0514-2 PMID: 31582815

25. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017; 7(11):170125. https://doi.org/10.1098/rsob.170125 PMID: 29093210

26. Fischer ST, Jiang Y, Broadaway KA, Conneely KN, Epstein MP. Powerful and robust cross-phenotype association test for case-parent trios. Genet Epidemiol. 2018; 42(5):447–458. https://doi.org/10.1002/gepi.22116 PMID: 29460449

27. Schaid DJ, Tong X, Batzler A, Sinnwell JP, Qing J, Biernacka JM. Multivariate generalized linear model for genetic pleiotropy. Biostatistics. 2019; 20(1):111–128. https://doi.org/10.1093/biostatistics/kxx067 PMID: 29267957

28. 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. PLoS Genet. 2020; 16 (12):e1009218. https://doi.org/10.1371/journal.pgen.1009218 PMID: 33290408

29. Schutte BC, Saal HM, Goudy S, Leslie E. *IRF6*-related disorders. In: GeneReviews. University of Washington, Seattle; 2014.

30. Yu Y, Zuo X, He M, Gao J, Fu Y, Qin C, et al. Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat Commun. 2017; 8(1):1–11. https://doi.org/10.1038/ncomms14364 PMID: 28232668

31. van Rooij IA, Ludwig KU, Welzenbach J, Ishorst N, Thonissen M, Galesloot TE, 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. 2019; 10(12):1023. https://doi.org/10.3390/genes10121023 PMID: 31817908

32. Khramtsova EA, Davis LK, Stranger BE. The role of sex in the genomics of human complex traits. Nat Rev Genet. 2019; 20(3):173–190. https://doi.org/10.1038/s41576-018-0083-1 PMID: 30581192

33. Harville EW, Wilcox AJ, Lie RT, Vindenes H, Åbyholm F. Cleft lip and palate versus cleft lip only: are they distinct defects? Am J Epidemiol. 2005; 162(5):448–453. https://doi.org/10.1093/aje/kwi214 PMID: 16076837

34. Ray D, Pankow JS, Basu S. USAT: A unified score-based association test for multiple phenotype-genotype analysis. Genet Epidemiol. 2016; 40(1):20–34. https://doi.org/10.1002/gepi.21937 PMID: 26638693

35. Leslie EJ, Taub MA, Liu H, Steinberg KM, Koboldt DC, Zhang Q, 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. 2015; 96(3):397–411. https://doi.org/10.1016/j.ajhg.2015.01.004 PMID: 25704602

36. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, et al. Disruption of an AP-2*α* binding site in an *IRF6* enhancer is associated with cleft lip. Nat Genet. 2008; 40(11):1341–1347. https://doi.org/10.1038/ng.242 PMID: 18836445

37. Sylvester B, Brindopke F, Suzuki A, Giron M, Auslander A, Maas RL, et al. A Synonymous Exonic Splice Silencer Variant in *IRF6* as a Novel and Cryptic Cause of Non-Syndromic Cleft Lip and Palate. Genes. 2020; 11(8):903. https://doi.org/10.3390/genes11080903 PMID: 32784565

38. Mostowska A, Gaczkowska A, Żukowski K, Ludwig K, Hozyasz K, Wójcicki P, et al. Common variants in *DLG1* locus are associated with non-syndromic cleft lip with or without cleft palate. Clin Genet. 2018; 93 (4):784–793. https://doi.org/10.1111/cge.13141 PMID: 28926086

39. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse genome database (MGD) 2019. Nucleic Acids Res. 2019; 47(D1):D801–D806. https://doi.org/10.1093/nar/gky1056 PMID: 30407599

40. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 2018; 19(151):1–12. https://doi.org/10.1186/s13059-018-1519-9 PMID: 30286773

41. Zhang R, Cao L, Wang Y, Fang Y, Zhao L, Li W, et al. A unique methylation pattern co-segregates with neural tube defect statuses in Han Chinese pedigrees. Neurol Sci. 2017; 38(12):2153–2164. https://doi.org/10.1007/s10072-017-3132-1 PMID: 28980068

42. Wilson RD, Audibert F, Brock JA, Carroll J, Cartier L, Gagnon A, 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. 2015; 37(6):534–549. https://doi.org/10.1016/S1701-2163(15)30231-0 PMID: 26334606

43. Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. High-resolution epigenomic atlas of human embryonic craniofacial development. Cell Rep. 2018; 23(5):1581–1597. https://doi.org/10.1016/j.celrep.2018.03.129 PMID: 29719267

44. Cejudo-Martin P, Yuen A, Vlahovich N, Lock P, Courtneidge SA, Díaz B. Genetic disruption of the *sh3pxd2a* gene reveals an essential role in mouse development and the existence of a novel isoform of tks5. PLoS One. 2014; 9(9):e107674. https://doi.org/10.1371/journal.pone.0107674 PMID: 25259869

45. Warner DR, Mukhopadhyay P, Brock G, Webb CL, Michele Pisano M, Greene RM. Micro RNA expression profiling of the developing murine upper lip. Dev Growth Differ. 2014; 56(6):434–447. https://doi.org/10.1111/dgd.12140 PMID: 24849136

46. Duan J, Shi J, Fiorentino A, Leites C, Chen X, Moy W, 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. 2014; 95(6):744–753. https://doi.org/10.1016/j.ajhg.2014.11.001 PMID: 25434007

47. Bie Y, Zhang Z. *RAB8A* a new biomarker for endometrial cancer? World J Surg Oncol. 2014; 12(1):371. https://doi.org/10.1186/1477-7819-12-371 PMID: 25477298

48. Taub MA, Schwender H, Beaty TH, Louis TA, Ruczinski I. Incorporating genotype uncertainties into the genotypic TDT for main effects and gene-environment interactions. Genet Epidemiol. 2012; 36(3):225–234. https://doi.org/10.1002/gepi.21615 PMID: 22678881

49. Marazita ML, Lidral AC, Murray JC, Field LL, Maher BS, McHenry TG, 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. 2009; 68(3):151–170. https://doi.org/10.1159/000224636 PMID: 19521098

50. Ludwig K, Böhmer A, Rubini M, Mossey P, Herms S, Nowak S, et al. Strong association of variants around *FOXE1* and orofacial clefting. J Dent Res. 2014; 93(4):376–381. https://doi.org/10.1177/0022034514523987 PMID: 24563486

51. Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near *MAFB* and *ABCA4*. Nat Genet. 2010; 42(6):525–529. https://doi.org/10.1038/ng.580 PMID: 20436469

52. Liu H, Leslie EJ, Carlson JC, Beaty TH, Marazita ML, Lidral AC, et al. Identification of common non-coding variants at 1p22 that are functional for non-syndromic orofacial clefting. Nat Commun. 2017; 8:14759. https://doi.org/10.1038/ncomms14759 PMID: 28287101

53. Liu H, Busch T, Eliason S, Anand D, Bullard S, Gowans LJ, 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. 2017; 109(1):27–37. https://doi.org/10.1002/bdra.23596 PMID: 28029220

54. Chen WM, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. Am J Hum Genet. 2009; 85(3):364–376. https://doi.org/10.1016/j.ajhg.2009.08.003 PMID: 19732865

55. Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, 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. 2016; 25(13):2862–2872. https://doi.org/10.1093/hmg/ddw104 PMID: 27033726

56. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006; 38(2):209–213. https://doi.org/10.1038/ng1706 PMID: 16415888

57. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016; 48(10):1284–1287. https://doi.org/10.1038/ng.3656 PMID: 27571263

58. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet. 2013; 93(4):687–696. https://doi.org/10.1016/j.ajhg.2013.09.002 PMID: 24094745

59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015; 4(1). https://doi.org/10.1186/s13742-015-0047-8 PMID: 25722852

60. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10(2):giab008. https://doi.org/10.1093/gigascience/giab008 PMID: 33590861

61. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5(6):e1000529. https://doi.org/10.1371/journal.pgen.1000529 PMID: 19543373

62. Schwender H, Taub MA, Beaty TH, Marazita ML, Ruczinski I. Rapid Testing of SNPs and Gene-Environment Interactions in Case-Parent Trio Data Based on Exact Analytic Parameter Estimation. Biometrics. 2012; 68(3):766–773. https://doi.org/10.1111/j.1541-0420.2011.01713.x PMID: 22150644

63. Schwender H, Li Q, Neumann C, Taub MA, Younkin SG, Berger P, et al. Detecting disease variants in case-parent trio studies using the bioconductor software package `trio`. Genet Epidemiol. 2014; 38 (6):516–522. https://doi.org/10.1002/gepi.21836 PMID: 25048299

64. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.

65. Fallin D, Beaty T, Liang KY, Chen W. Power comparisons for genotypic vs. allelic TDT methods with >2 alleles. Genet Epidemiol. 2002; 23(4):458–461. https://doi.org/10.1002/gepi.10192 PMID: 12432512

66. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52(3):506–516. PMID: 8447318

67. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017; 8(1):1826. https://doi.org/10.1038/s41467-017-01261-5 PMID: 29184056

68. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26(18):2336–2337. https://doi.org/10.1093/bioinformatics/btq419 PMID: 20634204