Published in final edited form as:

Genet Epidemiol. 2022 April ; 46(3-4): 182–198. doi:10.1002/gepi.22447.

Genome-wide association study of multiethnic non-syndromic orofacial cleft families identifies novel loci specific to family and phenotypic subtypes

Nandita Mukhopadhyay[1,*], **Eleanor Feingold**[1,2,3], **Lina Moreno-Uribe**[4], **George Wehby**[5], **Luz Consuelo Valencia-Ramirez**[6], **Claudia P. Restrepo Muñeton**[6], **Carmencita Padilla**[7], **Frederic Deleyiannis**[8], **Kaare Christensen**[9], **Fernando A. Poletta**[10], **Ieda M Orioli**[11,12], **Jacqueline T. Hecht**[13], **Carmen J. Buxó**[14], **Azeez Butali**[15], **Wasiu L. Adeyemo**[16], **Alexandre R. Vieira**[1], **John R. Shaffer**[1,3], **Jeffrey C. Murray**[17], **Seth M. Weinberg**[1,3], **Elizabeth J. Leslie**[18,**], **Mary L. Marazita**[1,3,19,**]

[1] Center for Craniofacial and Dental Genetics, Department of Oral and Craniofacial Sciences, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, 15219 USA

[2] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

[3] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA

[4] Department of Orthodontics, & The Iowa Institute for Oral Health Research, College of Dentistry, University of Iowa, Iowa City, IA, USA

[5] Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA, USA

[6] Fundación Clínica Noel; Calle 14 # 43B – 146, Medellín, Antioquia, Colombia

[7] Department of Pediatrics, College of Medicine, Institute of Human Genetics, National Institutes of Health, University of the Philippines, Manila, the Philippines

[8] UCHealth Medical Group, Colorado Springs, CO. USA

[9] Unit of Epidemiology, Department of Public Health, University of Southern Denmark, Odense, Denmark

[10] CEMIC-CONICET: Center for Medical Education and Clinical Research, Buenos Aires, Argentina.

[11] Department of Genetics, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[12] Instituto Nacional de Genética Médica Populacional INAGEMP, Porto Alegre, Brazil.

[*] **Correspondence** Nandita Mukhopadhyay, nandita@pitt.edu.

[**] co-senior-authors

The authors have no conflicts of interest to declare.

[13] Department of Pediatrics, University of Texas Health Science Center at Houston, Houston, TX, USA

[14] Dental and Craniofacial Genomics Core, School of Dental Medicine, University of Puerto Rico, San Juan, Puerto Rico

[15] Department of Oral Pathology, Radiology and Medicine and Iowa Institute for Oral Health Research, College of Dentistry, University of Iowa, Iowa City, IA, USA

[16] Department of Oral and Maxillofacial Surgery, College of Medicine, University of Lagos, Lagos, Nigeria.

[17] Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, IA, USA

[18] Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

[19] Clinical and Translational Science, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Non-syndromic orofacial clefts (nsOFCs) are among the most common craniofacial birth defects worldwide, and known to exhibit phenotypic and genetic heterogeneity. Cleft lip plus cleft palate (CLP) and cleft lip only (CL) are commonly combined together as one phenotype (CL/P), separately from cleft palate alone. In comparison, our study analyzes CL and CLP separately. A sample of 2,218 CL and CLP cases, 4,537 unaffected relatives of cases, and 2,673 pure controls with no family history of OFC were selected from the Pittsburgh Orofacial Cleft (Pitt-OFC) multiethnic study. GWASs were run for seven specific phenotypes created based on the cleft type(s) observed within these families, as well as the combined CL/P phenotype. Five novel genome-wide significant associations, 3q29 (rs62284390), 5p13.2 (rs609659), 7q22.1 (rs6465810), 19p13.3 (rs628271) and 20q13.33 (rs2427238), and nine associations ($p \leq 1.0E-05$) within previously confirmed OFC loci - *PAX7*, *IRF6*, *FAM49A*, *DCAF4L2*, 8q24.21, *ARID3B*, *NTN1*, *TANC2* and the *WNT9B*:*WNT3* gene cluster – were observed. We also found that SNPs within a subset of the associated loci, both previously known and novel, differ substantially in terms of their effects across cleft- or family-specific phenotypes, indicating not only etiologic differences between CL and CLP, but also genetic heterogeneity within each of the two OFC subtypes.

INTRODUCTION

Orofacial clefts (OFCs) are among the most common birth defects worldwide. The physical health effects of OFCs pose social, emotional and financial burdens on affected individuals and their families (Berk & Marazita, 2002; Nidey et al., 2016; Wehby & Cassell, 2010), despite therapies such as surgical treatments, ongoing orthodontia, speech therapy etc. that are available to reduce these burdens. Similar to other birth-related malformations, there are disparities in access to the complex medical and surgical therapies for OFCs (Nidey & Wehby, 2019). A variety of studies have reported a reduced quality of life for children with OFC (Naros et al., 2018), as well as a higher risk of certain types of cancers in adulthood (Bille et al., 2005; Bui et al., 2018; Taioli et al., 2010). Thus, identifying etiologic factors responsible for OFCs is a very important tool for determining risk, designing prevention methods, and determining the extent of therapeutic and social support needed by individuals with OFCs and their families.

OFCs are heterogeneous with varying manifestations and severity but are typically categorized into three subtypes: cleft lip alone (CL), cleft palate alone (CP), and cleft lip plus cleft palate (CLP). These can be syndromic (i.e. part of a spectrum of multiple defects due to a single cause), but the majority, about 70% of CL with or without CP (CL/P) and 50% of CP, are non-syndromic (i.e. the only defect present without any other detectable cognitive or structural abnormality) (Dixon et al., 2011). Many of the genes responsible for Mendelian forms of syndromic OFCs have been identified (OMIM, https://www.omim.org/search/advanced/geneMap) as have some teratogenic causes. In contrast, our understanding of the genetic causes of non-syndromic OFCs (nsOFCs) remains incomplete due to the complex nature of these defects, despite studies over a number of years (Marazita & Leslie, 2016; Moreno Uribe & Marazita, In press). Not only are there differences in birth prevalence around the world with respect to any nsOFC, the prevalence of the various subtypes (CL, CLP, CP) also varies substantially, suggesting etiological differences in the genetic factors giving rise to these different forms of nsOFC. These differences likely reflect the fact that human craniofacial development is a multi-stage process involving complex interactions between genetic and environmental factors (Moreno Uribe & Marazita, In press).

Historically, CL and CLP have been treated as variants of the same defect based on embryological origins of the upper lip and secondary palate, with CLP being considered a more severe form of CL (Harville et al., 2005). Analysis of recurrence risk among siblings have shown that the cross-subtype recurrence risk ratio between CL and CLP is higher than between CP and either CL or CLP (Grosen et al., 2010), and analyzing the composite phenotype with lip involvement (CL/P) within association analyses have resulted in consistently stronger signals, than analyzing all three (CL, CLP, CP) as a combined phenotype. Therefore, CP has been treated as being genetically distinct from nsOFCs involving the lip. More recently, it has been shown that CL and CLP have shared and unique etiological factors, therefore, recent genetic studies have focused on investigating etiological differences between CL and CLP, including both candidate gene approaches (Carlson et al., 2019; Carlson et al., 2017) as well as genome-wide association study (GWAS) approaches (Huang et al., 2019; Moreno Uribe et al., 2017; Yu et al., 2017).

Our current study focuses on nsOFC and investigates whether CL is etiologically different from CLP by considering the types of clefts segregating within families. This family-type based approach was previously used for genome-wide linkage-analyses (Marazita et al., 2009), but has not been employed for GWASs. Following a methodology similar to the prior family-based analysis for partitioning families (Marazita et al., 2009), we created several GWAS samples and phenotypes, as defined in the Terminology section below, and described in detail in Methods. This approach stands in contrast to previous GWASs, including those utilizing Pittsburgh Orofacial Cleft Study (Pitt-OFC) participants (Leslie et al., 2017; Leslie et al., 2016) that have focused only on the *individual* subjects' cleft types (see e.g. table 04.02 in (Moreno Uribe & Marazita, In press)). The Pitt-OFC resource is a rich collection of nsOFC families across multiple racial/ethnic groups, including simplex, multiplex, and extended pedigrees (~12,000 participants) with precise and detailed information on the types of nsOFC observed within multiple generations of the relatives of the probands. This resource is therefore well suited to investigating differences between the genetic etiology of CL vs. that of CLP. Study samples were genotyped on a custom whole genome genotyping array, followed by imputation using the 1000 Genomes Project reference panel (phase 3). In our current study, we selected families containing one or more individuals affected with CL and/or CLP, excluding families with only CP.

Since the degree of OFC risk at certain susceptibility loci varies with ancestry (Mukhopadhyay et al., 2021), the effect of ancestry was incorporated into our analyses. The four ancestry groups used to classify study participants are AFR (African ancestry), ASIA (Asian ancestry), EUR (white, European ancestry) and CSA (Central and South American ancestry). EAF is used to denote the effect allele frequency within a specified subset of participants. LD r^2 is used to denote linkage disequilibrium between variants as observed within the POFC sample.

Terminology

Family types: Three non-overlapping types of families were considered, **CL** – all affected members have CL; **CLP** - all affected members have CLP; and **CL+CLP** - families containing CL as well as CLP affected members. Further, **CL+** designates the union of CL and CL+CLP families, **CLP+** designates the union of CLP and CL+CLP families, and **POFC** is used to designate the union of **CL**, **CLP** and **CL+CLP**.

Phenotypic subgroups: Eight phenotype analysis subgroups were defined based on the family types. Phenotypic subgroup designations list the analyzed OFC phenotypes with a subscript for the family type(s) included in each: $\mathbf{CL/P_{[POFC]}}$ is the full sample analyzed by assigning a positive affection status to both CL- and CLP-affected subjects. $\mathbf{CL_{[CL]}}$ is the GWAS sample and phenotype including pedigrees with only CL-affected (no CLP-affected) members, and $\mathbf{CLP_{[CLP]}}$ only CLP-affected (no CL-affected). $\mathbf{CL/P_{[CL+CLP]}}$ is the sample and phenotype consisting of pedigrees with both CL and CLP affecteds, assigning a positive affection status to both CL and CLP members. Similarly, $\mathbf{CL_{[CL+CLP]}}$ and $\mathbf{CLP_{[CL+CLP]}}$ are samples also consisting of pedigrees with both CL and CLP affecteds, but with only CL members set to affected (CLP members excluded), or only CLP members set to affected (CL members excluded) respectively. Finally, $\mathbf{CL_{[CL+]}}$ and $\mathbf{CLP_{[CLP+]}}$ are samples consisting of the CL+ or CLP+ family groups; respectively, but with only CL members set to affected (CLP members excluded), or only CLP members set to affected (CL members excluded). GWAS sample definition and phenotype assignment is described in detail in the Methods section. Table 1 lists selected prior studies of OFC types utilizing the Pitt-OFC subjects, that most closely resemble the subset and phenotypes analyzed in our study.

METHODS

Study sample

Our study sample consists of participants from the multiethnic Pittsburgh Orofacial Cleft study (Pitt-OFC) (Leslie et al., 2016), including a variety of pedigree structures and sizes, and including both simplex as well as multiplex families. Sample recruitment was carried out in accordance with ethics approval procedures at the University of Pittsburgh, the coordinating center for the Pitt-OFC study, as well as the respective institutions that contributed samples to the Pitt-OFC study. Genotyping was carried out at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University, on an Illumina chip for approximately 580,000 variants genome-wide as summarized previously (Leslie et al., 2017; Leslie et al., 2016), and available from dbGaP (**dbGaP Study Accession:** phs000774.v2.p1). The CIDR coordinating center at the University of Washington was also responsible for ensuring the quality of called genotypes. Subsequently, genotypes were imputed using the "1000 genome project phase 3" reference panel, at approximately 35,000,000 variants of the GrCH37 genome assembly. Genotyping, quality control, and imputation steps were previously described in detail in Leslie et al. (Leslie et al., 2016).

The full sample – POFC – utilized in our current study includes 2,218 individuals affected with CL or CLP, and 4,537 unaffected relatives from 1,939 families that contain members affected with CL and/or CLP. The types of OFCs present in a pedigree were obtained by direct participation by affected individuals and/or by a reported family history of OFCs. An additional 2,673 unaffected individuals from 1,474 families with no reported history of an OFC (referred to as Controls) are included in the association analysis. Participants from pedigrees containing individuals affected with a cleft palate only (CP), or having a reported family history of CP were excluded from this study.

Definition of subtypes

Several subsets were created from the POFC sample based on the types of OFCs reported within pedigrees, as follows. First, the pedigrees were partitioned into three non-overlapping subsets, (i) [CL]: pedigrees that contain individuals affected with CL only, but not members affected with CLP, (ii) [CLP]: pedigrees that contain individuals affected with CLP but not members affected with CL only, and (iii) [CL+CLP]: pedigrees containing some members affected with CL only as well as some members affected with CLP. The partitioning of pedigrees into these three subsets used all available phenotypic and relationship information, including phenotypic information from pedigree members who were not genotyped. Two additional subsets were then defined, (iv) [CL+], all pedigrees with any CL-affected member, i.e. the union of [CL] and [CL+CLP], and (v) [CLP+], all pedigrees with any CLP-affected member, i.e. the union of [CLP] and [CL+CLP]. The [CL+] and [CLP+] subsets are not disjoint, i.e. they both contain subjects from [CL+CLP] pedigrees.

Eight GWAS phenotypic subtypes were then defined for these five subsets of pedigrees for running genome-wide association analysis, and affection statuses assigned to pedigree members belonging to each of the eight phenotypic subtypes as described below. The 2,673 Controls were included in each of the GWASs.

A. **CL/P$_{\textbf{[POFC]}}$** – Within the full POFC sample, participants with either a CL, or CLP were set to affected, participants without any OFC were set to unaffected.

B. **CL$_{\textbf{[CL]}}$** –Within the [CL] pedigrees - group (i) above, participants with CL were set to affected, and those without CL were set to unaffected.

C. **CLP**$_{\mathbf{[CLP]}}$ – Within the [CLP] group of pedigrees – group (ii), participants with CLP were set to affected, and those without CLP were set to unaffected.

D. **CL/P**$_{\mathbf{[CL+CLP]}}$ – within the [CL+CLP] group of pedigrees – group (iii), participants with either CL or CLP were set to affected, and those without OFCs were set to unaffected.

E. **CL**$_{\mathbf{[CL+CLP]}}$ –Within [CL+CLP] pedigrees – group (iii), participants with a CL only were set to affected, those with CLP were set to unknown (thereby excluding them from GWAS), and those without OFCs were set to unaffected.

F. **CLP**$_{\mathbf{[CL+CLP]}}$ –Within [CL+CLP] pedigrees – group (iii), pedigree members with a CLP were set to affected, those with CL only were set to unknown (thereby excluding them from GWAS), and those without OFCs were set to unaffected.

G. **CL**$_{\mathbf{[CL+]}}$ – Within the [CL+] – group (iv) pedigrees, participants with CL only were set to affected, those with CLP were set an unknown affection status (thereby excluding them from GWAS), and those without any OFC were set to unaffected.

H. **CLP**$_{\mathbf{[CLP+]}}$ – Within the [CLP+] pedigrees – group (v), participants affected with CLP were set to affected, those with CL only were set to unknown (thereby excluding them from GWAS), and those without any OFC were set to unaffected.

Figure 1 shows the partitioning of POFC pedigrees into the eight phenotypic subsets and phenotype definitions within each of these phenotypic subsets that were used to run separate GWASs. For illustration purposes, each subtype is depicted as simple nuclear pedigree structures with three offspring, two of which are affected with CL or CLP, although a wide variety of family types are represented in this study. Simplex and multi-generation pedigrees were handled following the same procedure for grouping into subtypes. In addition to the type of pedigrees included in each subset, Figure 1 also depicts affected and unaffected members, as well as those assigned an unknown affection status, thereby excluding these members from the corresponding GWAS.

Genome wide association

We have shown previously that the degree of OFC risk at certain susceptibility loci varies with ancestry of the sample participants (Mukhopadhyay et al., 2021). In order to control for this variance, we first classified subjects into four different genetically defined ancestry groups using the principal component analysis-based classification defined in a previous study using POFC subjects (Leslie et al., 2016). For each of the eight GWAS phenotypic samples defined above and shown in Figure 1, we first analyzed each ancestry group separately, then combined the association outcomes using meta-analysis. The four ancestry-based groups were: AFR (participants of African origin), ASIA (participants of Asian origin), EUR (those of European white origin), and CSA (participants of Central and Southern American origin). Table 3 shows the breakdown of the analysis sample by ancestry, pedigree type, and affection status.

Individual GWASs were run using the mixed-model association program, GENESIS (Gogarten et al., 2019). GENESIS uses a genetic relationship matrix (GRM) estimated from the observed genotype data to account for population structure and familial relatedness, therefore, it is not necessary to correct for population admixture using ancestry PCs. The use of a GRM is necessary to account for population admixture within our ancestry-based subsets, which, in turn is due to the varying geographical origin of participants in each of these subsets (see Supplementary Table S2 for a breakdown by recruitment site). The genetic relationship matrix also provides an estimate of the polygenic variance component. Significance of association is based on the score test, comparing the maximum likelihood of disease outcomes conditional on observed genotypes at each variant to the maximum likelihood of the unconditional polygenic model. GENESIS reports approximate effect sizes in the form of betas, i.e. the log-likelihood ratio of the conditional and unconditional model) and standard error of the effect size. In this study, the effect allele is fixed across all GWASs as the minor allele at each variant identified in the combined POFC sample.

Ancestry-specific GWASs were then meta-analyzed for each of the eight GWAS phenotypes using the inverse-variance method implemented in PLINK (Chang et al., 2015). The reported odds ratios from PLINK were converted to log-scale effect sizes, to conform to the GENESIS reported effects. The 95% confidence intervals of betas were calculated under the assumption that the meta-analysis p-values are distributed normally. All four ancestry-groups were meta-analyzed for the $CL_{[CL+]}$ and $CLP_{[CLP+]}$ subtypes. There are no AFR pedigrees containing both CL and CLP affected members, therefore, meta-analysis was conducted excluding the African samples (AFR) for the five family-subtypes ($CL_{[CL]}$, $CL_{[CL+CLP]}$, $CLP_{[CLP]}$, $CLP_{[CL+CLP]}$ and $CL/P_{[CL+CLP]}$).

Variant selection

Genotyped and imputed variants that passed quality control, and had minor allele frequencies of 2% or more within their respective GWAS sample subsets were used to run association. The observed minor allele frequencies of reported loci were checked against values obtained from the gnomAD database (Karczewski et al., 2019) to guard against imputation inaccuracy.

Identification of novel associations

For each genome-wide meta-analysis, variants showing association p-values below 1.0E-06 were selected for further investigation, and grouped into association peaks measuring 1MB or less. We then checked for overlap between our associations peaks with the 29 genomic regions listed as harboring known OFC genes by Beaty et al. (Beaty et al., 2016) as well as associated regions reported by six recently published OFC GWAS studies. The six recent GWASs include (1) combined meta-analysis of parent-offspring trio and case-control cohorts from the current Pitt-OFC multiethnic study sample (Leslie et al., 2016), (2) meta-analysis of the cohorts used in (1) with another OFC sample consisting of European and Asian participants (Leslie et al., 2017), (3) GWAS of cleft lip with cleft palate in Han Chinese samples (Yu et al., 2017), (4) GWAS of cleft lip only and cleft palate only in Han Chinese (Huang et al., 2019), (5) GWAS of cleft lip with or without cleft palate in Dutch and Belgian participants (van Rooij et al., 2019) and (6) GWAS of sub-Saharan African participants from Nigeria, Ghana, Ethiopia and the Republic of Congo (Butali et al., 2019).

For each OFC gene, we checked if any our 1 MB association peaks overlapped with the span of the gene, as determined by its start and end transcription sites. The base pair positions for start and end transcription sites were obtained from the UCSC genome browser (https://genome.ucsc.edu/index.html) mapped to the February 2009 (GRCh37) assembly. For the 8q24.21 locus, which is a gene desert, we checked whether any of our associated SNPs were located in the 8q24.21 chromosome band. The distance between variants published by the six recent GWASs and our variants with p-values below 1.0E-06 were similarly measured, and a positive overlap reported if this distance was less than 500 Kb.

Comparison of association outcomes between subtypes

Within each peak region the variant with the smallest meta-analysis association p-value observed for each of the eight subtypes were selected and their effect sizes compared. Effect size of each variant is represented by the beta coefficient of the SNP main effect under an additive model of inheritance, setting the minor allele (based on the entire POFC study sample) as the effect allele. Effect size and magnitude were compared across subtypes for the variants selected for each subtype to determine whether the 95% confidence intervals of effect size estimates overlapped. Next, LD r^2 between selected variants at each locus was calculated using the PLINK program and the set of genotyped founders in the full POFC sample, irrespective of their OFC status. Finally, the observed effect allele frequency (EAF) within cases from the two GWASs were examined to assess whether these differed significant between cleft subtypes. We have previously shown that ancestry impacts association to CL/P in our POFC sample (Mukhopadhyay et al., 2021); therefore, we examined the subtype-specific effect sizes within each ancestry group to assess whether the differences observed were similar to the those observed for the meta-analysis. EAFs within cases were also compared across the eight phenotypic subtypes within each ancestry group in addition to the cases pooled across ancestry groups for each phenotypic subset. In our study, we did not carry out a statistical test (e.g. Cochran's Q statistic) to compare association outcomes from the OFC subtypes, as the unaffected relatives of OFC subjects and subjects from control families were used in the GWAS of more than one subtype; therefore, we relied mainly on qualitative evaluation of differences in the association outcomes.

RESULTS

In our study, GWASs of eight separate phenotypes were run on eight corresponding phenotypic subsets created by grouping the POFC pedigrees based on the type of OFCs (CL and/or CLP) observed within those pedigrees. The full sample was analyzed for the CL/P phenotype (CL/$P_{[POFC]}$), and seven other phenotype/family groups, $CL_{[CL]}$, $CLP_{[CLP]}$, CL/$P_{[CL+CLP]}$, $CL_{[CL+CLP]}$, $CLP_{[CL+CLP]}$, $CL_{[CL+]}$ and $CLP_{[CLP+]}$ were defined, and analyzed using GWASs. For each phenotype, pedigrees were further grouped according to their population ancestry groups, and GWASs run separately within each group. Subsequently, association outcomes for the ancestry groups were meta-analyzed to determine association for each of the eight phenotypic subsets. The procedure followed for creating and analyzing the eight phenotypic subgroups is described in the Methods section. Genome-wide meta-analysis resulted in several significant and suggestive associations, both at previously reported OFC loci, and five novel regions.

Significant and suggestive loci identified by meta-analysis

Meta-analysis over the ancestry groups for each of the eight phenotypes resulted in fourteen unique loci of interest. These included five novel loci with genome-wide Bonferroni significant meta-analysis p-values ($p < 5.0e-08$) and an additional nine known OFC loci with p-values below 1.0E-06. Table 2 lists the most significant meta-analysis p-value, effect size (expressed as betas), 95% CI of the effect size, and the variant positions that showed significant ($p < 5.0e-08$) or suggestive ($p < 1.0e-05$) associations. Supplementary Table S1 provides more detailed information for all variant positions corresponding to the p-values shown in Table 2, such as RS numbers, base pair positions, and effect allele frequencies (EAFs) within the affected subjects included for GWAS of that phenotype.

The five novel associations observed are: (i) 3q29, most significantly associated with the $CL_{[CL+CLP]}$ subtype, (ii) 5q13.2, most significantly associated with the $CL_{[CL+]}$ subtype, (iii) 7q22.1 showing the strongest association with the CLP $_{[CL+CLP]}$ subtype, (iv) 19p13.3 also showing the strongest association with the $CLP_{[CL+CLP]}$ subtype, and (v) 20q13.3, associated with the $CL_{[CL]}$ subtype.

The known OFC loci recapitulated here include the genes *PAX7, IRF6, FAM49A, DCAF4L2, ARID3B, NTN1, WNT9B:WNT3, TANC2*, and the 8q24.21 locus. Among these, *PAX7, FAM49A*, *DCAF4L2*, *ARID3B*, and *WNT9B:WNT3* are associated with both CL and CLP. The *IRF6* locus is the most strongly associated with the $CL_{[POFC]}$ subtype, *TANC2* with the $CL_{[CL]}$ subtype, and *NTN1* with $CLP_{[CLP]}$ subtype. The 8q24.21 locus has traditionally been treated as a single locus, however, the prior CL/P GWAS study using samples from Pitt-OFC reported two distinct peak regions with genome-wide significant association p-values (Leslie et al. (Leslie et al., 2016)). In the current study, we also observed two distinct peak regions at this locus. Both peaks are most strongly associated with the $CL/P_{[POFC]}$ subtype.

Identification of loci associated with specific cleft and/or family subtypes

Based on the strength of association and location of the most significant variants across subtypes, six previously reported OFC loci, *PAX7*, *FAM49A*, *DCAF4L2*, the 8q24.21 locus, *ARID3B*, *WNT9B:WNT3* and a novel locus 7q22.1 appear to be associated with both CL and CLP, i.e., the $CL/P_{[POFC]}$ meta p-values were the most significant at these loci with subtypes represented by the larger samples - $CLP_{[CLP+]}$ and $CLP_{[CLP]}$ - produced more significant association p-values as compared to the subtypes with smaller samples. The remaining nine loci produced more significant p-values within a cleft or a family subtype. We hypothesized that the differences in p-values could be the result of the sample size differences between phenotypic subtypes. We therefore compared the estimated meta-analysis effect sizes of the associated variants within each of 15 peak regions identified above obtained for the eight phenotypes. This was done to verify whether the degree of risk for developing an OFC differed by OFC type and/or family type.

Table 2 lists the estimated beta coefficients and 95% confidence intervals for the top associated variant at each locus and for each subtype GWAS. The comparison showed statistically significant differences between the meta-analysis beta coefficients between subtypes at five of the associated loci, both between cleft subtypes (i.e. CL[$_{CL+]}$ vs. CLP$_{[CLP+]}$) and between family subtypes (i.e. CL$_{[CL]}$, CLP$_{[CLP]}$, CL$_{[CL+CLP]}$ and CLP$_{[CL+CLP]}$). A comparison of the ancestry-specific beta coefficients also showed variation similar to the meta-analysis effect sizes. A comparison of the frequency of the effect allele within affected individuals included in the phenotypic subsets showed that subtype-specific variants occurred at varying frequencies between subgroups. Overall, case allele frequencies were observed to differ between subtypes if effect sizes varied between subtypes, and vice versa.

Three of the loci considered as being associated with a specific subtype, are presented in figures 2–4 below. Figure 2 shows the *IRF6* locus; Figure 3 and Figure 4 show two interesting novel loci - 20q13.33 and 3q29; each containing multiple variants associated with genome-wide significant and/or suggestive p-values. These three figures illustrate that subtype-specific differences in strength of association mostly correspond to effect size differences, and also to differences in frequency of the effect allele amongst affected subjects (referred to as case EAFs) belonging to these subtypes. Differences in effect sizes and case EAFs that are observed at the meta-analysis level are also seen within ancestry groups, especially the two largest ones - CSA and EUR.

In each figure, the top panel (a) shows a regional Manhattan plot with the most significant association per subtype – the top associations are labelled in order of their genomic position. Panel (b) in each figure shows the LD pattern of variants with p-value below 0.001 as that locus - LD r^2 values above 0.2 shaded as indicated, and top associations labelled as in panel (a). Overall, LD patterns between top associations from the subtypes are as expected, i.e. LD is high between subtype-specific associations that are in close proximity, low (> 0.2) otherwise. Panel (c) shows the effect size estimates (beta coefficient and 95% CI) for the labelled associations for all subtypes – effect size estimates of significant and suggestive associations are identified in the forest plot, and the lead SNP name outlined. Panel (d) compares ancestry-subgroup specific effect sizes for either the two cleft subtypes (CL$_{[CL+]}$ and CLP$_{[CLP+]}$), or the four family subtypes (CL$_{[CL]}$, CLP$_{[CLP]}$, CL$_{[CL+CLP]}$, CLP$_{[CL+CLP]}$) at the lead SNP depending on which comparisons indicated subtype specificity. Panel (e) compares effect allele frequency within affected subjects in each subtype to that of controls at the lead SNP by ancestry. The observed variation in effect sizes across subtypes corresponds to differences in case EAFs, i.e. case EAFs within subtypes differ from one another, if the effect sizes are different, with a single exception – the 5q13.2 locus, which is further explored in the next section.

1. Loci specific to the CL cleft-subtype—The novel locus at **5q13.2**, and the known **1q32.2 (*IRF6*)** locus show the most significant association for the CL$_{[POFC]}$ cleft subtype. Figure 2 shows the *IRF6* locus in detail: the regional Manhattan plot (Figure 1a) shows six distinct variants (labelled A-F) with the most significant p-values from the subtype meta-analyses. The top association for $CL_{[CL+]}$ coincides with the top $CL_{[CL+CLP]}$ variant (SNP D: rs67652997 in Fig 2c), although the latter shows lower significance, and the top associations for $CLP_{[CLP+]}$ and $CLP_{[CLP]}$ also coincide (SNP B: rs2076149). LD between variants with significance p-values (below 0.001) is shown for the 209.92–209.98 KB region spanning five of these variants (A-E); the top $CL_{[CL]}$ association is not shown - it is in low LD with the rest of the top associations.

The largest CL effect size is observed for the $CL_{[CL+]}$ subtype, as can be seen in Figure 2c for *IRF6*. The $CL_{[CL+]}$ subtype's effect sizes at the lead SNP rs609659, as well as nearby variants in LD with the lead SNP is distinctly larger in magnitude than for the $CLP_{[CLP+]}$ subtype. Effect sizes for the $CL_{[CL]}$ and $CL_{[CL+CLP]}$ family-based subtypes are also larger than the $CLP_{[CLP]}$ and $CLP_{[CL+CLP]}$ effect sizes, while $CL_{[CL]}$ and $CL_{[CL+CLP]}$ effect sizes are not statistically different. These loci show stronger association to CL, attributable to both the $CL_{[CL]}$ and $CL_{[CL+CLP]}$ family subtypes. Within the *IRF6* gene, the lead variant is observed to have a protective effect on CL risk and observed at a lower frequency than the non-effect allele within cases in EUR and CSA. Within ASIA and AFR, effect sizes appear to be similar between $CL_{[CL+]}$ and $CLP_{[CLP+]}$. At the 5q13.2 locus, the ancestry subgroup-specific effect sizes are consistent with the meta-analysis effect sizes within the ASIA, EUR and CSA subgroups, i.e. $CL_{[CL+]}$ effect sizes are larger in magnitude than $CLP_{[CLP+]}$. Beta coefficients overlap within the AFR subgroup. The EAF within $CL_{[CL+]}$ affecteds of all ancestries pooled is not different from the EAF in $CLP_{[CLP+]}$ cases, unlike variants within the other subtype-specific loci. However, this appears to be due to EAF differences across ancestry groups: in AFR, the $CL_{[CL+]}$ EAF is smaller than the $CLP_{[CLP+]}$, while the reverse is true in ASIA, EUR and CSA (supplement Figure S1).

2. Loci specific to the $CL_{[CL]}$ family-subtype—At two peak regions, the novel locus at 20q13.33, and 17q23.2;q23.3 (*TANC2*), the $CL_{[CL]}$ meta-analysis p-value is the most significant, and the $CL_{[CL]}$ meta-analysis effect sizes are much larger than the other family-type based subsets. Notably, the $CL_{[CL+]}$ effect size is not different from the $CLP_{[CLP+]}$ subtype. Figure 3 highlights the main association outcomes at the 20q13.33 locus. As seen in Figure 3d, the variation in beta estimates within the CSA and EUR subgroups correspond to the variation observed within the overall meta-analysis beta estimates, and the lead variant for $CL_{[CL]}$ shows a positive effect size (beta), while other effect sizes are close to zero. The effect allele was not observed in $CL_{[CL]}$ families from ASIA, and AFR was excluded from the family-subtype comparison (Figure 3e). At the other locus showing association within the $CL_{[CL]}$ subtype - *TANC2*, effect size differences were observed in the EUR and CSA group, with differences observed in the ASIA group. Further, within the CSA group, the $CL_{[CL]}$ subtype showed a positive effect whereas the $CL_{[CL+CLP]}$ subtype showed a negative effect, which was not the case for EUR. EAFs within the affecteds were consistently highest in the $CL_{[CL]}$ subtype sample than the other family-subtypes, and the effect allele is least frequent in ASIA (Supplement Figure S2).

3. 3q29 locus specific to $CL_{[CL+CLP]}$ family-subtype—The **3q29 novel locus** is more strongly associated with the $CL_{[CL+CLP]}$ subtype than any other subtype (Figure 4).

There is low LD between SNPs associated with different subtypes as seen in Figure 4b. The $CL_{[CL+CLP]}$ subtype's effect size is much larger than that of other subtypes also resulting in a significant difference between the $CL_{[CL+]}$ subtype's effect size and the $CLP_{[CLP+]}$ subset's effect size (Figure 4c and 4d). The **3q29** locus is another instance where ancestry plays a role. The elevated beta in $CL_{[CL+CLP]}$ is due to samples of EUR ancestry, and the corresponding EAF in the EUR subgroup is also much higher than EAFs of other family subtypes (Figure 4e). Effect size variation is not observed in CSA, which is consistent with similar case EAFs in CSA, and the effect allele is very rarely observed in ASIA. When effect sizes from the ancestry-based subgroups are examined, the difference between $CL_{[CL]}$ and $CL_{[CL+CLP]}$ effect sizes is observed in the EUR subgroup, but not in ASIA and CSA.

4. **Locus specific to $CLP_{[CL+CLP]}$ family-subtype—**The **19p13.3 peak** includes a single Bonferroni-significant association at SNP rs628271; with no other neighboring variants reaching a suggestive level of significance, this may not be a reliable association. Even so, interestingly the effect size of this variant for the $CLP_{[CL+CLP]}$ subtype is larger than all the other family-based subtypes. The $CL_{[CL+]}$ subtype effect size is similar to the $CLP_{[CLP+]}$ effect size. This difference is observed in CSA and EUR, but not in ASIA.

5. **Loci with no variation in subtype-specific effect sizes:** At the following loci, the subtype-specific effect sizes are similar in magnitude and direction to those from the other subtypes, indicating that that these loci affect the risk of both CL and CLP to a similar extent regardless of family classification: 1p36.13 (*PAX7*), 2p24.2–24.3 (*FAM49A*), 7q22.1 - novel locus, 8q21.3 (*DC4FL2*), both peaks within 8q24.1, 15q24.1;q24.2 (*ARID3B*), 17p13.1 (*NTN1*), and 17q21.31;q21.32 (*WNT9B;WNT3*). At these loci, larger samples yielded more significant association p-values.

DISCUSSION

For the five novel loci observed in our study, a bioinformatics search yielded interesting, but not conclusive indication of their roles in the development of OFCs. The lead variant within 5q13.2 is in close proximity to the *TMEM1* gene, and the lead variant within the 20q13.33 locus is intronic to the *CDH4* gene; both *TMEM1* and *CDH4* are involved in the Wnt signaling pathway, known to be involved in the development of OFCs. The lead variant in our 3q29 locus is located approximately 1 MB downstream of the *DLG1* gene, reported as being associated with CL/P in a recent study of CL/P on a Polish population (Mostowska et al., 2018). In our study, however, we observed only weak association to variants within the *DLG1* gene. The other three loci contain craniofacial super-enhancer regions. The top associations in the 7q22.1 locus are intronic to the *COL26A1* and *RANBP3* genes, both reported as having a blood phenotype (UCSC genome browser, https://genome.ucsc.edu/index.html). It is interesting to note that the previously reported genome-wide linkage and targeted region study of Pitt-OFC pedigree subsets based on cleft types (Marazita et al., 2009) reported two regions – 9q21.33 and 14q21.3 – that were associated at a suggestive level of significance in our study, although the current associations do not lie within the fine-mapped regions analyzed in the former study. Further in-depth study of these novel loci including fine-mapping and functional analysis need to be conducted to identify their role in the formation of OFCs.

The analysis of CL and CLP as a single phenotype (CL/P) in the [CL+CLP] families did not produce unique associations, as would be expected if these families were segregating for genes that cause a continuum of the CL/P phenotype. This lack of association may further support the hypothesis that CL/P is not a single phenotype etiologically. Further, we hypothesize that our family subtype-based analyses show evidence of genetic heterogeneity even within the cleft subtypes CL and CLP themselves. For example, association of CL to *TANC2* is much stronger in the [CL] families than in the [CL+CLP] families, while the reverse is true at the 3q29 locus. Also notably, our study outcomes show consistently stronger and more reliable associations for the CL-based subtypes (5 previously known and novel loci) as compared to the CLP-based subtypes (a single novel locus), although the sample sizes for the CLP-based subtypes are larger. Our study results recapitulated the association of *IRF6* with CL (Rahimov et al., 2008). We thus hypothesize that CL is genetically more homogeneous than CLP. A possible alternative to genetic heterogeneity would be phenotypic heterogeneity: there exists diagnostic uncertainty with the palate phenotype, it is sometimes left undiagnosed, or, in some cases, the presence of submucous CP along with CL is not categorized as CLP. However, Pitt-OFC subjects were thoroughly examined for submucous CP and VPI, so this would be unlikely to have happened on large enough scale to impact our analysis outcomes.

This study makes an important contribution to the study of heterogeneity between OFC types using a study design where both the individuals as well as the family's OFC types are incorporated. The idea that genetically related individuals also tend to have the same type of OFC more often than different types of OFCs, has been rarely utilized in running GWASs of OFC subtypes. Individual level phenotypic heterogeneity in terms of laterality and gender-effects has been analyzed to investigate genetic heterogeneity in previous studies on the Pitt-OFC subjects (Curtis, Chang, Lee, et al., 2021; Curtis, Chang, Sun, et al., 2021). However, these variables were not incorporated in our study – as our sample is already sub-divided into smaller samples based on the OFC subtypes observed in families, further subsetting by gender and/or laterality would lead to subsets too small for running GWAS. Our study found strong evidence for genetic heterogeneity of OFCs – four out of the five novel associations detected in this study were obtained through the family-type subset analysis, rather than CL and CLP subtype analysis, although the latter samples are larger. We conclude that our phenotype definition based on pedigree information resulted in the creation of more genetically homogeneous subsets, identifying genomic regions that are specific to an OFC subtype. Our study provides a methodology for incorporating the proband's relatives' cleft types within the GWAS framework, and the observed outcomes provide valuable insight into etiological differences between OFC subtypes.

The knowledge of etiological differences between OFC subtypes is a key step in the prevention, treatment and management of OFCs. Although genetic findings of OFC GWASs are yet to be used clinically, we hypothesize a few such possible applications in the future. Variants that are involved in detoxification or nutritional uptake may be used as preventative measures, and those that play a role in wound-healing may be used to determine the best course of treatment of OFC affected individuals e.g. *IRF6* has been reported to interact with maternal exposures such as tobacco smoke and multivitamin supplementation (Wu et al., 2010), and to play a role in wound healing such as this study (Carlson et al., 2017) among several others. Another possible application could be cancer screening – the risk of certain types of cancer is increased in individuals with OFCs, mutations in genomic regions associated with OFCs may therefore be used for cancer screening as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors wish to thank the participant families worldwide, without whom this research would not have been possible. Special thanks to Dr. Eduardo Castilla (deceases), Dr. Juan C. Mereb, Dr. Andrew Czeizel, and to the devoted staff at the many recruitment sites. This work was supported by grants from the National Institutes of Health including: X01-HG007485 [MLM], R01-DE016148 [MLM, SMW], U01-DE024425 [MLM], R37-DE008559 [JCM, MLM], R01-DE009886 [MLM], R21-DE016930 [MLM], R01-DE014667 [LMM], R21-DE016930 [MLM], R01-DE012472 [MLM], R01-DE011931 [JTH], U01-DD000295 [GLW], R00-DE025060 [EJL], R01- DE028342 [EJL], R01- DE28300 [AB], K99/R00DE024571 [CJB], S21MD001830 [CJB], and U54GM133807 [CJB]. Genotyping and data cleaning were provided via an NIH contract to the Johns Hopkins Center for Inherited Disease Research: HHSN268201200008I. Additional support provided by: an intramural grant from the Research Institute of the Children's Hospital of Colorado [FWD]; operating costs support in the Philippines was provided by the Institute of Human Genetics, National Institutes of Health, University of the Philippines, Manila [CP]; grants through FAPERJ [IMO].

BIBLIOGRAPHY

Beaty TH, Marazita ML, & Leslie EJ (2016). Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. F1000Res, 5, 2800. 10.12688/f1000research.9503.1 [PubMed: 27990279]

Berk NW, & Marazita ML (2002). The Costs of Cleft Lip and Palate: Personal and Societal Implications. In Wyszynski DF(Ed.), Cleft Lip and Palate: From Origin to Treatment. Oxford University Press.

Bille C, Winther JF, Bautz A, Murray JC, Olsen J, & Christensen K (2005). Cancer risk in persons with oral cleft--a population-based study of 8,093 cases. Am J Epidemiol, 161(11), 1047–1055. 10.1093/aje/kwi132 [PubMed: 15901625]

Bui AH, Ayub A, Ahmed MK, Taioli E, & Taub PJ (2018). Association Between Cleft Lip and/or Cleft Palate and Family History of Cancer: A Case-Control Study. Ann Plast Surg, 80(4 Suppl 4), S178–s181. 10.1097/sap.0000000000001331 [PubMed: 29389703]

Butali A, Mossey PA, Adeyemo WL, Eshete MA, Gowans LJJ, Busch TD, Jain D, Yu W, Huan L, Laurie CA, Laurie CC, Nelson S, Li M, Sanchez-Lara PA, Magee WP, Magee KS, Auslander A, Brindopke F, Kay DM, Caggana M, Romitti PA, Mills JL, Audu R, Onwuamah C, Oseni GO, Owais A, James O, Olaitan PB, Aregbesola BS, Braimah RO, Oginni FO, Oladele AO, Bello SA, Rhodes J, Shiang R, Donkor P, Obiri-Yeboah S, Arthur FKN, Twumasi P, Agbenorku P, Plange-Rhule G, Oti AA, Ogunlewe OM, Oladega AA, Adekunle AA, Erinoso AO, Adamson OO, Elufowoju AA, Ayelomi OI, Hailu T, Hailu A, Demissie Y, Derebew M, Eliason S, Romero-Bustillous M, Lo C, Park J, Desai S, Mohammed M, Abate F, Abdur-Rahman LO, Anand D, Saadi I, Oladugba AV, Lachke SA, Amendt BA, Rotimi CN, Marazita ML, Cornell RA, Murray JC, & Adeyemo AA (2019). Genomic analyses in African populations identify novel risk loci for cleft palate. Hum Mol Genet, 28(6), 1038–1051. 10.1093/hmg/ddy402 [PubMed: 30452639]

Carlson JC, Anand D, Butali A, Buxo CJ, Christensen K, Deleyiannis F, Hecht JT, Moreno LM, Orioli IM, Padilla C, Shaffer JR, Vieira AR, Wehby GL, Weinberg SM, Murray JC, Beaty TH, Saadi I, Lachke SA, Marazita ML, Feingold E, & Leslie EJ (2019). A systematic genetic analysis and visualization of phenotypic heterogeneity among orofacial cleft GWAS signals. Genet Epidemiol, 43(6), 704–716. 10.1002/gepi.22214 [PubMed: 31172578]

Carlson JC, Taub MA, Feingold E, Beaty TH, Murray JC, Marazita ML, & Leslie EJ (2017). Identifying Genetic Sources of Phenotypic Heterogeneity in Orofacial Clefts by Targeted Sequencing. Birth Defects Res, 109(13), 1030–1038. 10.1002/bdr2.23605 [PubMed: 28762674]

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

Curtis SW, Chang D, Lee MK, Shaffer JR, Indencleef K, Epstein MP, Cutler DJ, Murray JC, Feingold E, Beaty TH, Claes P, Weinberg SM, Marazita ML, Carlson JC, & Leslie EJ (2021). The PAX1 locus at 20p11 is a potential genetic modifier for bilateral cleft lip. HGG Adv, 2(2). 10.1016/j.xhgg.2021.100025

Curtis SW, Chang D, Sun MR, Epstein MP, Murray JC, Feingold E, Beaty TH, Weinberg SM, Marazita ML, Lipinski RJ, Carlson JC, & Leslie EJ (2021). FAT4 identified as a potential modifier of orofacial cleft laterality. Genet Epidemiol, 45(7), 721–735. 10.1002/gepi.22420 [PubMed: 34130359]

Dixon MJ, Marazita ML, Beaty TH, & Murray JC (2011). Cleft lip and palate: understanding genetic and environmental influences. Nat Rev Genet, 12(3), 167–178. 10.1038/nrg2933 [PubMed: 21331089]

Gogarten S, Sofer T, Chen H, Yu C, Brody J, Thornton T, Rice K, & Conomos M (2019). Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics. 10.1093/bioinformatics/btz567

Grosen D, Chevrier C, Skytthe A, Bille C, Mølsted K, Sivertsen A, Murray JC, & Christensen K (2010). A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. J Med Genet, 47(3), 162–168. 10.1136/jmg.2009.069385 [PubMed: 19752161]

Harville EW, Wilcox AJ, Lie RT, Vindenes H, & Abyholm F (2005). Cleft lip and palate versus cleft lip only: are they distinct defects? Am J Epidemiol, 162(5), 448–453. 10.1093/aje/kwi214 [PubMed: 16076837]

Huang L, Jia Z, Shi Y, Du Q, Shi J, Wang Z, Mou Y, Wang Q, Zhang B, Wang Q, Ma S, Lin H, Duan S, Yin B, Lin Y, Wang Y, Jiang D, Hao F, Zhang L, Wang H, Jiang S, Xu H, Yang C, Li C, Li J, Shi B, & Yang Z (2019). Genetic factors define CPO and CLO subtypes of nonsyndromicorofacial cleft. PLoS Genet, 15(10), e1008357. 10.1371/journal.pgen.1008357 [PubMed: 31609978]

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Vallabh Minikel E, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, & MacArthur DG (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. 531210. 10.1101/531210 %J bioRxiv

Leslie EJ, Carlson JC, Shaffer JR, Butali A, Buxo CJ, Castilla EE, Christensen K, Deleyiannis FW, Leigh Field L, Hecht JT, Moreno L, Orioli IM, Padilla C, Vieira AR, Wehby GL, Feingold E, Weinberg SM, Murray JC, Beaty TH, & Marazita ML (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. Hum Genet, 136(3), 275–286. 10.1007/s00439-016-1754-7 [PubMed: 28054174]

Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, Jain D, Laurie CC, Doheny KF, McHenry T, Resick J, Sanchez C, Jacobs J, Emanuele B, Vieira AR, Neiswanger K, Lidral AC, Valencia-Ramirez LC, Lopez-Palacio AM, Valencia DR, Arcos-Burgos M, Czeizel AE, Field LL, Padilla CD, Cutiongco-de la Paz EM, Deleyiannis F, Christensen K, Munger RG, Lie RT, Wilcox A, Romitti PA, Castilla EE, Mereb JC, Poletta FA, Orioli IM, Carvalho FM, Hecht JT, Blanton SH, Buxo CJ, Butali A, Mossey PA, Adeyemo WL, James O, Braimah RO, Aregbesola BS, Eshete MA, Abate F, Koruyucu M, Seymen F, Ma L, de Salamanca JE, Weinberg SM, Moreno L, Murray JC, & Marazita ML (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24.2, 17q23 and 19q13. Hum Mol Genet, 25(13), 2862–2872. 10.1093/hmg/ddw104 [PubMed: 27033726]

Marazita ML, & Leslie EJ (2016). Genetics of Nonsyndromic Clefting. In Losee J & Kirschner R (Eds.), Comprehensive Cleft Care (Second ed., pp. 207–224). CRC Press.

Marazita ML, Lidral AC, Murray JC, Field LL, Maher BS, Goldstein McHenry T, Cooper ME, Govil M, Daack-Hirsch S, Riley B, Jugessur A, Felix T, Morene L, Mansilla MA, Vieira AR, Doheny K, Pugh E, Valencia-Ramirez C, & Arcos-Burgos M (2009). Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. Hum Hered, 68(3), 151–170. 10.1159/000224636 [PubMed: 19521098]

Moreno Uribe LM, Fomina T, Munger RG, Romitti PA, Jenkins MM, Gjessing HK, Gjerdevik M, Christensen K, Wilcox AJ, Murray JC, Lie RT, & Wehby GL (2017). A Population-Based Study of Effects of Genetic Loci on Orofacial Clefts. J Dent Res, 96(11), 1322–1329. 10.1177/0022034517716914 [PubMed: 28662356]

Moreno Uribe LM, & Marazita ML (In press). Epidemiology, Etiology and Genetics of Orofacial Clefting. In Shetye P& Gibson TL (Eds.), Cleft and Craniofacial Orthodontics. Wiley.

Mostowska A, Gaczkowska A, Żukowski K, Ludwig KU, Hozyasz KK, Wójcicki P, Mangold E, Böhmer AC, Heilmann-Heimbach S, Knapp M, Zadurska M, Biedziak B, Budner M, Lasota A, Daktera-Micker A, & Jagodziński PP (2018). Common variants in DLG1 locus are associated with non-syndromic cleft lip with or without cleft palate. Clin Genet, 93(4), 784–793. 10.1111/cge.13141 [PubMed: 28926086]

Mukhopadhyay N, Feingold E, Moreno-Uribe L, Wehby G, Valencia-Ramirez LC, Muneton CPR, Padilla C, Deleyiannis F, Christensen K, Poletta FA, Orioli IM, Hecht JT, Buxo CJ, Butali A, Adeyemo WL, Vieira AR, Shaffer JR, Murray JC, Weinberg SM, Leslie EJ, & Marazita ML (2021). Genome-Wide Association Study of Non-syndromic Orofacial Clefts in a Multiethnic Sample of Families and Controls Identifies Novel Regions. Front Cell Dev Biol, 9, 621482. 10.3389/fcell.2021.621482 [PubMed: 33898419]

Naros A, Brocks A, Kluba S, Reinert S, & Krimmel M (2018). Health-related quality of life in cleft lip and/or palate patients - A cross-sectional study from preschool age until adolescence. J Craniomaxillofac Surg, 46(10), 1758–1763. 10.1016/j.jcms.2018.07.004 [PubMed: 30054220]

Nidey N, Moreno Uribe LM, Marazita MM, & Wehby GL (2016). Psychosocial well-being of parents of children with oral clefts. Child Care Health Dev, 42(1), 42–50. 10.1111/cch.12276 [PubMed: 26302988]

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

Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, & Murray JC (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet, 40(11), 1341–1347. 10.1038/ng.242 [PubMed: 18836445]

Taioli E, Ragin C, Robertson L, Linkov F, Thurman NE, & Vieira AR (2010). Cleft lip and palate in family members of cancer survivors. Cancer Invest, 28(9), 958–962. 10.3109/07357907.2010.483510 [PubMed: 20569073]

van Rooij IA, Ludwig KU, Welzenbach J, Ishorst N, Thonissen M, Galesloot TE, Ongkosuwito E, Bergé SJ, Aldhorae K, Rojas-Martinez A, Kiemeney LA, Vermeesch JR, Brunner H, Roeleveld N, Devriendt K, Dormaar T, Hens G, Knapp M, Carels C, & Mangold E (2019). 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 (Basel), 10(12). 10.3390/genes10121023

Wehby GL, & Cassell CH (2010). The impact of orofacial clefts on quality of life and healthcare use and costs. Oral Dis, 16(1), 3–10. 10.1111/j.1601-0825.2009.01588.x [PubMed: 19656316]

Wu T, Liang KY, Hetmanski JB, Ruczinski I, Fallin MD, Ingersoll RG, Wang H, Huang S, Ye X, Wu-Chou YH, Chen PK, Jabs EW, Shi B, Redett R, Scott AF, & Beaty TH (2010). Evidence of gene-environment interaction for the IRF6 gene and maternal multivitamin supplementation in controlling the risk of cleft lip with/without cleft palate. Hum Genet, 128(4), 401–410. 10.1007/s00439-010-0863-y [PubMed: 20652317]

Yu Y, Zuo X, He M, Gao J, Fu Y, Qin C, Meng L, Wang W, Song Y, Cheng Y, Zhou F, Chen G, Zheng X, Wang X, Liang B, Zhu Z, Fu X, Sheng Y, Hao J, Liu Z, Yan H, Mangold E, Ruczinski I, Liu J, Marazita ML, Ludwig KU, Beaty TH, Zhang X, Sun L, & Bian Z (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat Commun, 8, 14364. 10.1038/ncomms14364 [PubMed: 28232668]

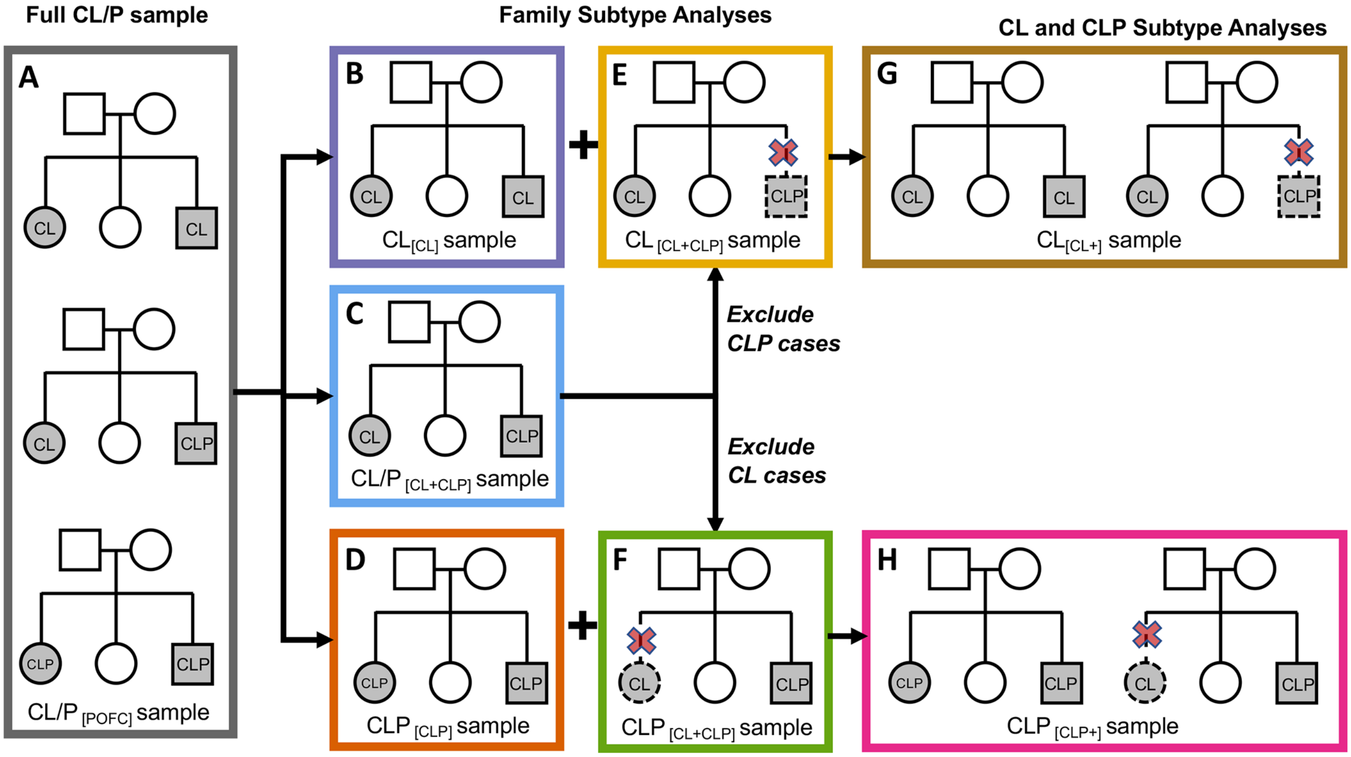

Figure 1:

Creation of analytical subsets and phenotype assignment for GWAS Each colored rectangle is a GWAS phenotypic subset; included pedigree type(s) shown for each subset; shaded squares and circles indicate participants with an OFC; shaded circles and squares with solid outlines indicate **affected** subjects; unshaded squares and circles with solid outlines represent **unaffected** subjects; circles and squares with dotted outlines represent pedigree members **excluded** from the GWAS; designations for OFC phenotype analysis subgroups including a subscript for the family type(s) are:

(A) **CL/P$_{\mathbf{[POFC]}}$**: full set of [CL], [CLP] and [CL+CLP] pedigrees, CL and CLP members set to affected;

(B) **CL$_{\mathbf{[CL]}}$**: in [CL] pedigrees, CL members are set to affected;

(C) **CLP$_{\mathbf{[CLP]}}$**, in [CLP] pedigrees CLP members are set to affected;

(D) **CL/P$_{\mathbf{[CL+CLP]}}$**, in [CL+CLP] pedigrees, CL and CLP members are set to affected;

(E) **CL$_{\mathbf{[CL+CLP]}}$**, in [CL+CLP] pedigrees, CL members set to affected, CLP members excluded;

(F) **CLP**$_{[CL+CLP]}$, in [CL+CLP] pedigrees, CLP members are set to affected and CL members excluded;

(G) **CL$_{\mathbf{[CL+]}}$**, in [CL+] pedigrees (i.e. [CL] plus [CL+CLP] pedigrees), CL members are set to affected and CLP members excluded;

(H) **CLP$_{\mathbf{[CLP+]}}$**, in [CLP+] pedigrees (i.e. [CLP] plus [CL+CLP] pedigrees), CLP members are set to affected and CL members excluded.

Affected sibships are shown as examples – data includes other pedigree types including multi-generational pedigrees.

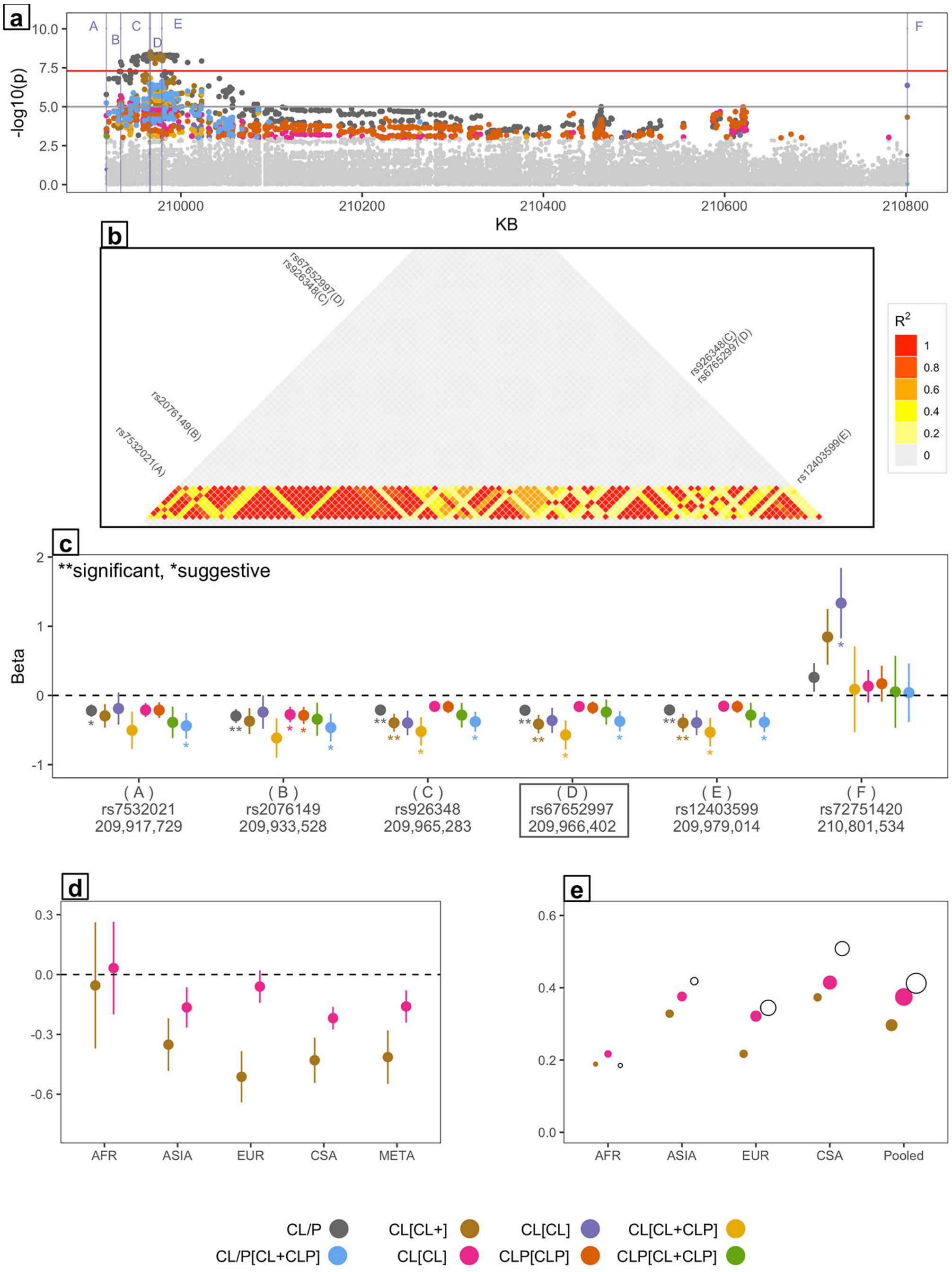

Figure 2. *IRF6* locus specific to CL$_{[CL+]}$ subtype

(a) Regional Manhattan plot consisting of six distinct variants (A-F) with the most significant p-value from each subtype; (b) LD r^2 values > 0.2 between variants (A-E) with p-value below 0.001, variant F is in a separate LD block from the A-E; (c) beta coefficient and 95% CI for variants A-F, D: lead variant at this locus, ** significant and * suggestive associations; (d) effect sizes and (e) effect allele frequency within affected subjects for cleft subtypes CL$_{[CL+]}$ vs. CLP$_{[CLP+]}$ by ancestry-subgroup.

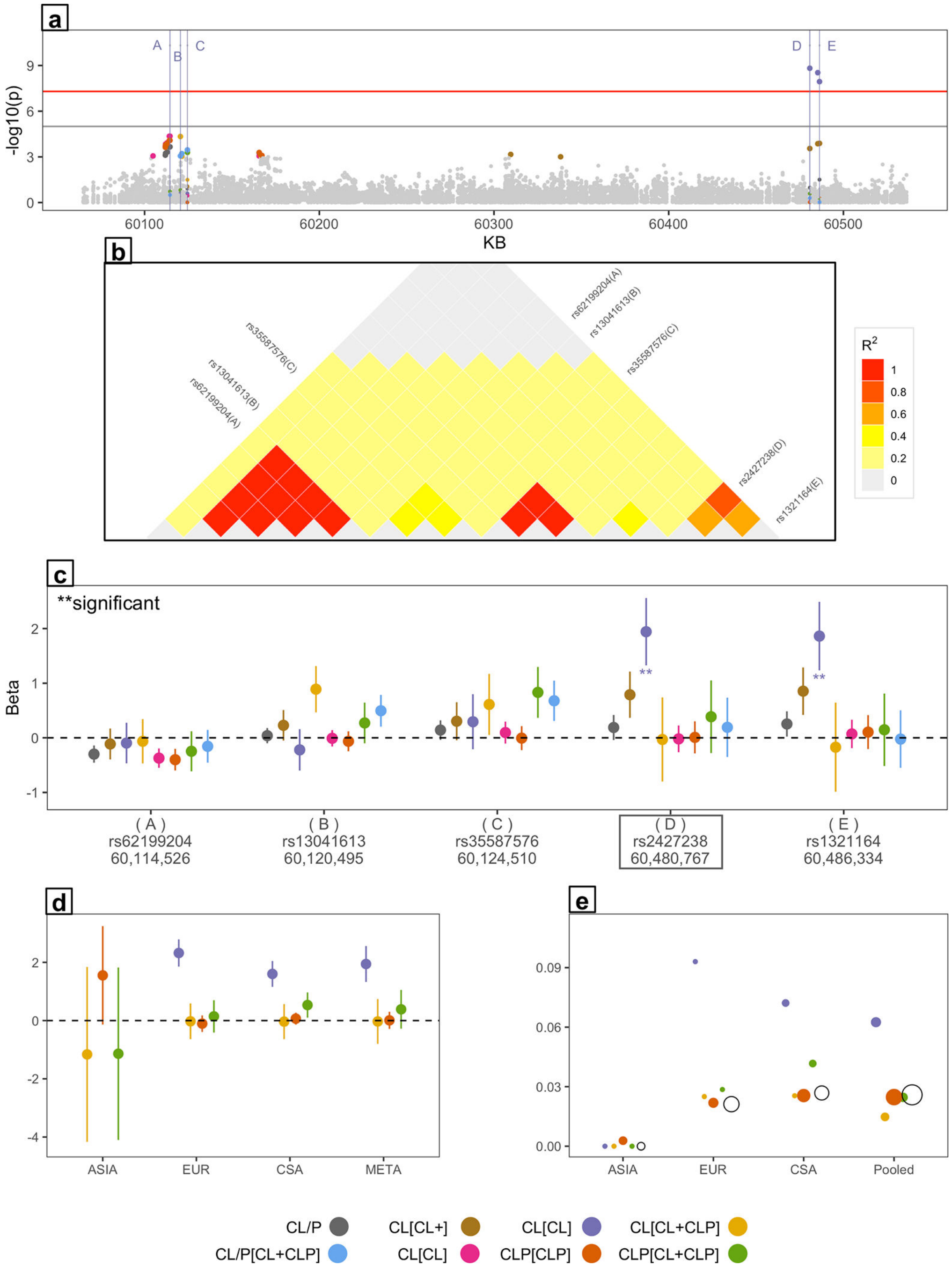

Figure 3. 20q13.3 novel locus specific to $CL_{[CL]}$ subtype

(a) regional Manhattan plot consisting of five distinct variants (A-E) with the most significant p-value from each subtype; (b) LD r^2 values > 0.2 between variants (A-E) with p-value below 0.001; (c) beta coefficient and 95% CI for variants A-E, D: lead variant at this locus, ** significant associations; (d) effect sizes and (e) effect allele frequency within affected subjects for family subtypes $CL_{[CL]}$, $CL_{[CL+CLP]}$, $CLP_{[CLP]}$ and $CLP_{[CL+CLP]}$ by ancestry-subgroup.

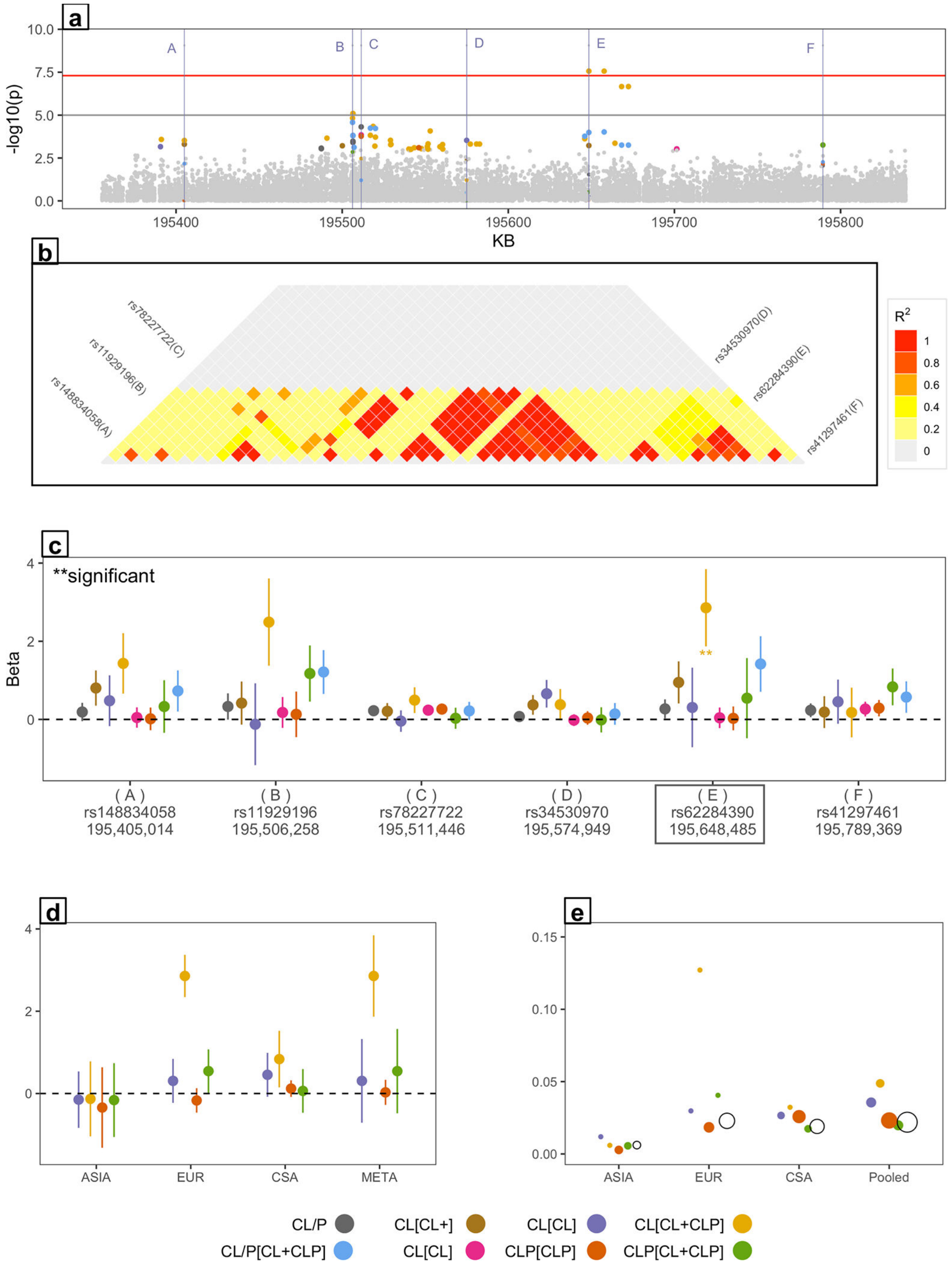

Figure 4. 3q29 novel locus specific to $CL_{[CL+CLP]}$ subtype

(a) regional Manhattan plot consisting of six distinct variants (A-F) with the most significant p-value from each subtype; (b) LD r^2 values > 0.2 between variants (A-F) with p-value below 0.001; (c) beta coefficient and 95% CI for variants A-F, E: lead variant at this locus, ** significant associations; (d) effect sizes, and (e) effect allele frequency within affected subjects for family subtypes $CL_{[CL]}$, $CL_{[CL+CLP]}$, $CLP_{[CLP]}$ and $CLP_{[CL+CLP]}$ by ancestry-subgroup.

Table 1.

Comparison of previous published analyses on Pitt-OFC

Prior Study	Study type/goal	Approach	Correspondence to current study subsets
Marazita et al. 2009 (Marazita et al., 2009)	Genome-wide linkage, fine-mapping	Parametric linkage (HLOD) and FBAT	$CL/P_{[POFC]}$, $CL_{[CL]}$, $CLP_{[CLP]}$
Leslie et al. 2017 (Leslie et al., 2017)	GWAS	TDT, case-control association and meta-analysis	$CL/P_{[POFC]}$
Carlson et al. 2019 (Carlson et al., 2019)	Heterogeneity within OFC in targeted gene regions	GWAS, meta-analysis, and heterogeneity Q-statistic with permutation testing for significance	$CL_{[CL+]}$, $CLP_{[CLP+]}$

Table 2.

Counts of pedigrees and participants by GWAS, ancestry and affection status

GWAS	CSA			EUR			ASIA			AFR		
	[†]Ped	[††]Case	UFM	Ped	Case	UFM	Ped	Case	UFM	Ped	Case	UFM
CL/P	954	1,050	1,889	511	569	1,373	321	445	1,081	153	154	194
CL$_{[CL+]}$	219	166	523	181	153	586	164	171	620	57	59	60
CLP$_{[CLP+]}$	847	884	1,667	427	416	1,123	260	274	890	96	95	134
CL$_{[CL]}$	102	101	222	84	90	250	61	85	191	57	59	60
CLP$_{[CLP]}$	725	762	1,336	328	339	787	157	184	461	96	95	134
CL$_{[CL+CLP]}$	117	65	301	97	63	336	103	86	429	0	0	0
CLP$_{[CL+CLP]}$	122	122	301	99	77	336	103	90	429	0	0	0
CL/P$_{[CL+CLP]}$	127	187	301	99	140	336	103	176	429	0	0	0
	Ped		**Ctrl**	**Ped**		**Ctrl**	**Ped**		**Ctrl**	**Ped**		**Ctrl**
CONTROL [†††]	478		1,098	759		1,330	163		165	74		80

Note:

[†] Ped=number of pedigrees, Case=number of affected individuals

[††] UFM=unaffected family member related to a case

[†††] the CONTROL subset consists of individuals/families with no known personal nor family history of OFCs, and are utilized in each GWAS – the number of CONTROL subjects are listed in the Ctrl columns to complete counts of unaffected GWAS subjects.

Table 3.

Loci with meta-analysis p-value ≤ 1.0E-06 in one or more GWASs

Locus	CL/P$_{\text{[POFC]}}$	CL$_{\text{[CL+]}}$	CL$_{\text{[CL]}}$	CL$_{\text{[CL+CLP]}}$	CLP$_{\text{[CLP+]}}$	CLP$_{\text{[CLP]}}$	CLP$_{\text{[CL+CLP]}}$	CL/P$_{\text{[CL+CLP]}}$
1p36.13 (*PAX7*)	**rs9439714 (C)** **5.9E-09, 0.24 ± 0.08**	1.6E-04, 0.87 ± 0.45	8.0E-05, 1.23 ± 0.61	1.0E-04, 1.16 ± 0.58	rs56675509 (C) 5.3E-08, 0.26 ± 0.09	rs11583072 (T) 1.7E-08, 0.28 ± 0.1	7.1E-04, 1.11 ± 0.64	2.3E-04, 0.28 ± 0.15
1q32.2 (*IRF6*)	rs926348 4.2E-09, −0.21 ± 0.07	**rs67652997 (A)** **3.0E-09, −0.41 ± 0.13**	rs72751420 (C) 4.3E-07, 1.33 ± 0.51	rs67652997 (A) 1.5E-07, −0.57 ± 0.21	2.0E-06, −0.28 ± 0.11	5.4E-06, −0.29 ± 0.12	7.9E-04, −0.39 ± 0.23	rs12403599 (C) 2.5E-07, −0.39 ± 0.14
2p24.2-p24.3 (*FAM49A*)	**rs7552 (G)** **1.2E-07, 0.19 ± 0.07**	2.5E-04, 0.25 ± 0.14	9.0E-04, 0.32 ± 0.19	6.0E-05, 1.05 ± 0.51	6.4E-06, 0.19 ± 0.08	2.6E-04, 0.17 ± 0.09	5.7E-05, 0.37 ± 0.18	7.3E-05, 0.30 ± 0.15
3q29†	4.8E-05, 0.22 ± 0.11	5.0E-04, 0.80 ± 0.45	3.0–04, 0.66 ± 0.35	**rs62284390 (T)** **2.7E-08, 2.86 ± 0.99**	1.4E-04, 0.24 ± 0.12	1.7E-04, 0.26 ± 0.14	5.5E-04, 0.83 ± 0.47	2.6E-05, 1.21 ± 0.56
5q13.2†	1.1E-03, −0.12 ± 0.07	**rs609659 (G)** **4.6E-08, −0.39 ± 0.14**	1.6E-06, −0.47 ± 0.19	6.9E-04, −0.50 ± 0.29	1.2E-04, 0.32 ± 0.16	2.1E-03, 0.29 ± 0.18	4.1E-03, 0.28 ± 0.19	3.4E-03, −0.47 ± 0.32
7q22.1†	4.9E-04, 0.16 ± 0.09	1.8E-03, 0.32 ± 0.2	5.2E-03, 1.33 ± 0.93	1.8E-04, 1.17 ± 0.61	1.7E-03, 0.20 ± 0.13	2.2E-03, 0.27 ± 0.17	**rs6465810 (C)** **1.2E-08, 1.17 ± 0.4**	rs6465810 (C) 5.7E-07, 0.78 ± 0.3
8q21.3 (*DCAF4L2*)	**rs12543318 (C)** **5.4E-10, 0.22 ± 0.07**	3.6E-05, 0.27 ± 0.13	7.6E-04, −0.33 ± 0.19	1.1E-03, 0.44 ± 0.26	rs12543318 (C) 2.9E-08, 0.22 ± 0.08	4.5E-06, 0.21 ± 0.09	8.8E-06, 0.39 ± 0.17	6.2E-06, 0.32 ± 0.14
8q24.21(*p-ter*)	**rs7839784 (T)** **2.2E-14, 0.34 ± 0.08**	rs55768865 (G) 5.2E-09, 0.88 ± 0.29	rs55768865 (G) 5.1E-07, 1.04 ± 0.4	1.9E-05, 1.04 ± 0.47	rs5894949 (A) 2.2E-11, 0.33 ± 0.09	rs55768865 (G) 5.6E-10, 0.56 ± 0.17	2.8E-06, 0.54 ± 0.22	rs55768865 (G) 1.9E-07, 0.84 ± 0.31
8q24.21(*q-ter*)	**rs72728755 (A)** **3.1E-32, 0.58 ± 0.09**	rs72728755 (A) 6.4E-13, 0.72 ± 0.19	rs112704402 (A) 1.3E-08, 0.74 ± 0.25	rs72728755 (A) 1.5E-08, 0.91 ± 0.31	rs72728755 (A) 1.5E-26, 0.59 ± 0.1	rs72728755 (A) 8.6E-20, 0.58 ± 0.12	rs72728755 (A) 1.7E-14, 1.04 ± 0.26	rs72728755 (A) 1.2E-16, 0.89 ± 0.2
15q24.2-q24.1 (*ARID3B*)	**rs58691516 (CT)** **6.7E-07, −0.20 ± 0.08**	9.0E-06, −0.40 ± 0.17	1.2E-03, 0.35 ± 0.21	3.2E-04, −0.39 ± 0.21	4.0E-05, −0.19 ± 0.09	5.7E-06, −0.24 ± 0.1	1.2E-04, −0.48 ± 0.24	9.2E-04, −0.30 ± 0.18

Locus	CL/P$_{[POFC]}$	CL$_{[CL+]}$	CL$_{[CL]}$	CL$_{[CL+CLP]}$	CLP$_{[CLP+]}$	CLP$_{[CLP]}$	CLP$_{[CL+CLP]}$	CL/P$_{[CL+CLP]}$
17p13.1 (*NTN1*)	rs12944377 (C) 6.3E-09, −0.22 ± 0.07	1.5E-04, −0.29 ± 0.15	1.6E-04, −0.39 ± 0.2	3.3E-03, 0.32 ± 0.21	rs12944377 (C) 1.7E-09, −0.26 ± 0.08	**rs12944377 (C) 5.8E-11, −0.31 ± 0.09**	5.8E-04, 0.75 ± 0.43	1.6E-04, 0.29 ± 0.15
17q21.31-q21.32 (*WNT9B:WNT3*)	**rs7216951 (T) 3.0E-07, −0.22 ± 0.08**	9.2E-04, −0.40 ± 0.24	3.5E-03, −0.31 ± 0.21	5.4E-04, −0.60 ± 0.34	rs7216951 (T) 4.8E-07, −0.25 ± 0.09	3.6E-06, −0.26 ± 0.11	7.3E-04, −0.52 ± 0.3	1.4E-03, 0.23 ± 0.14
17q23.3-q23.2 (*TANC2*)	2.6E-06, −0.22 ± 0.09	1.2E-04, 0.30 ± 0.15	**rs17683292 (C) 1.0E-07, 0.56 ± 0.2**	1.3E-04, −0.42 ± 0.22	1.7E-05, −0.23 ± 0.1	6.4E-06, −0.23 ± 0.12	7.1E-04, −0.73 ± 0.42	1.8E-03, −0.58 ± 0.36
19p13.3[†]	1.5E-03, 0.33 ± 0.2	6.4E-04, 0.25 ± 0.14	1.1E-03, 0.76 ± 0.46	2.9E-03, 0.39 ± 0.26	1.4E-04, 0.44 ± 0.23	2.4E-03, 0.41 ± 0.27	**rs628271 (C) 1.9E-08, 1.68 ± 0.58**	1.7E-04, 0.92 ± 0.48
20q13.33[†]	2.2E-04, −0.30 ± 0.16	1.3E-04, 0.85 ± 0.43	**rs2427238 (G) 1.5E-09, 1.94 ± 0.62**	4.6E-05, 0.89 ± 0.42	4.4E-05, −0.37 ± 0.18	8.2E-05, −0.40 ± 0.2	5.0E-04, 0.83 ± 0.47	3.4E-04, 0.68 ± 0.37

Note: For each locus and GWAS, meta-analysis p-value, beta estimate and its 95% CI are shown; RS numbers and their effect alleles (in parentheses) are shown for suggestive and significant associations - SNPs with the most significant p values at a locus may differ across the GWASs

[†] novel loci; p-values ≤ 5.0E-08 highlighted in dark green and p-values ≤ 1.0E-06 in light green; smallest p-value across subtypes highlighted in bold; two distinct association peaks in 8q24.21 locus listed separately.