Identification and comparison of imputed and genotyped variants for genome-wide association study of orofacial cleft case-parent trios

By

Wanying Zhang, MD

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Epidemiology

Baltimore, Maryland
April 2019

Abstract:

Background: Orofacial clefts (OFCs) – cleft lip with/without cleft palate (CL/P) and cleft palate (CP) – are the most common craniofacial malformations among newborns. Both CL/P and CP show strong familial aggregation resulting in high estimated heritability. Previously identified genetic risk factors account for about a quarter of the estimated total heritability of risk to OFCs, indicating additional genetic risk loci remain to be identified. The aim of this thesis is to update imputed genotypes generated from a genome-wide marker panel and use both observed and imputed genetic variants to identify the genetic risk factors for OFCs in a case-parent trio study of OFC.

Methods: We imputed genotypes on case-parent trios from the Genes and Environment Association (GENEVA) consortium using the Michigan Imputation Server, and then conducted genome-wide association analysis to identify genetic variants associated with risk of CL/P and CP separately. For each cleft subtype, we performed genotypic transmission disequilibrium test (gTDT) using the *trio* R package on common single nucleotide polymorphic (SNP) markers (i.e. those with a minor allele frequency [MAF] $\geq$ 5%) in all the trios together, and then stratified by ethnicity (Asian and European sub-groups).

Results: We identified two genes not previously reported as associated with risk to CL/P - 18q12 (*TTR*) and 4q22 (*GRID2*). The most significant SNP in the region of *TTR* (rs1375445) reached genome-wide significance in the combined set of all trios ($p = 4.33 \times 10^{-8}$) with RR=1.35 [95%CI: (1.21, 1.51)], despite not achieving this level of significance in either the European sub-group ($p = 2.94 \times 10^{-5}$) or Asian sub-group ($p = 5.52 \times 10^{-5}$) separately. However, the most significant SNP of *GRID2* (rs1471079) reached genome-wide significance only in the Asian sub-group ($p = 1.82 \times 10^{-7}$) with estimated RR = 0.70 [95%CI:

(0.60, 0.80)]. Both of these imputed SNPs have high imputation accuracy (rs1375445 R^2 = 0.96; rs1471079 R^2 = 0.97). Additionally, for CL/P, we replicated significant association of 8 regions identified in previous studies of these case-parent trios, including 8q24 (recognized as a gene desert), 1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*). The most significant SNPs in six of these regions were imputed. The most significant SNP (rs17242358) in the 8q24 region showed genome-wide significance ($p = 1.75 \times 10^{-16}$) in the combined set of all trios. This imputed SNP showed over-transmission of A allele (over G allele) with estimated RR = 2.09 [95%CI: (1.76, 2.49)]. This imputed SNP achieved quite different levels of significance in the European ($p = 7.11 \times 10^{-14}$) and Asian sub-groups ($p = 7.3 \times 10^{-4}$) primarily because the MAF differed across the two sub-groups (MAF = 23% in Europeans and 2% in Asians). We did not detect any genome-wide significant locus for the OFC subtype CP.

Conclusions: Our findings confirm the complex genetic architecture and the heterogeneity of genes influencing risk to OFCs. We replicated most previously reported genetic risk factors for these GENEVA case-parent trios. We also identified two new genetic risk factors for CL/P that require further investigation. Stratification by racial groups helped detect OFC risk loci specific to certain groups. In addition, imputation helped improve the statistical power to detect genetic risk factors for OFC.

Keyword:

orofacial clefts, cleft lip, cleft palate, genome-wide association study, imputation, 8q24

Thesis Committee

Dr. Debashree Ray (Advisor)

Dr. Terri H. Beaty (Co-advisor)

Dr. Margaret A. Taub

Thesis Readers

Dr. Debashree Ray

Dr. Margaret A. Taub

Table of Contents

List of Tables

List of Figures

Background

Background on orofacial clefts

Orofacial clefts (OFCs) are the most common craniofacial malformations among newborns and include three anatomically distinct malformations: cleft lip (CL), cleft lip with cleft palate (CLP) and cleft palate (CP). OFCs can occur as an isolated malformation, with another structural malformation (i.e. the infant has 2 or more congenital anomalies) or as part of a recognized malformation syndrome, some of which are Mendelian syndromes directly attributable to mutations in a single gene [1]. Cleft lip with/without cleft palate (CL/P) and CP are distinct with respect to their different embryologic origins, where the outer face develops before the inner palate closes [2]. CL/P results from the lack of fusion of lateral nasal, median nasal and maxillary mesodermal processes, whereas CP occurs due to a failure of the palatal shelves to fuse about week 12 of embryologic development [2]. Previous studies suggest 70 percent of CL/P cases and 50 percent of CP cases occur as isolated, non-syndromic malformations [3].

The genetic etiology, recurrence risks and surgical treatments also vary between CL/P and CP. Genes controlling cell patterning, cell proliferation and differentiation of the midface are all good candidate genes for OFC malformations [4]. Non-syndromic OFCs are regarded as genetically complex and heterogeneous, influenced by multiple genes, recognized environmental risk factors (e.g. maternal smoking and alcohol consumption) plus the potential for both gene-gene [5] and gene-environment [6] interactions. Over two dozen candidate genes have been identified as contributing to risk of OFC by genome-wide association studies, but these recognized genes can only explain about a quarter of the observed heritability of OFC [7].

Children with OFCs usually require extensive multidisciplinary care, which includes feeding assistance, plastic surgery, otolaryngology care, developmental follow-up, and speech therapy throughout childhood. OFC patients have increased mortality due to difficulties in breastfeeding [8] and usually suffer from social discrimination during their lifetime [9]. Due to the high prevalence and huge financial and psychological burden of OFCs, understanding the etiology of OFCs and improving the health of newborns are important public health goals.

Prevalence and Epidemiology

As reported by the National Birth Defects Prevention Network (NBDPN) for 2007 to 2011, the estimated birth prevalence of all OFCs (i.e. CL/P and CP combined, including isolated and syndromic cases) for 29 states in the US was 14.5/10,000 live births: the birth prevalence for CL/P was 8.7/10,000 live births, whereas CP occurred in 5.9/10,000 live births. Among all newborns with CL/P, approximately one-third presented with CL alone and two thirds presented with CLP [10].

Worldwide, the birth prevalence of OFC varies considerably by race and ethnicity, with lowest rates of CL/P in populations of African ancestry (10.2/10,000 live births), highest in American Indians (20.5/10,000 live births), and intermediate in other racial groups (e.g. Non-Hispanic White 15.4/10,000 live births, Asian 13.2/10,000 live births) [10]. However, these differences in birth prevalence worldwide represent true differences and differences in case ascertainment and surveillance methods [11]. For example, the prevalence of CL/P among newborns in Japan (20.0/10,000 live births) is almost twice the birth prevalence reported in the United States and Canada [12]. Differences in birth prevalence of CP also have been reported but are complicated by the difficulty in diagnosing CP during the newborn period

[10]. Moreover, CL/P tends to affect more males than females, whereas CP affects more females than males [13].

Environmental risk factors

Several medications have been reported as teratogens for midfacial development. Antiseizure agents such as phenytoin and topiramate are commonly administrated drugs recognized to increase risk of OFC [14]. Additionally, folate is an essential component in the process of DNA methylation, and maternal folate levels can influence risk to OFC. Deficiency in folate, which may result from the folic acid antagonist methotrexate, contributes to risk of multiple birth defects including OFCs [15]. Several previous studies have shown maternal smoking [16] and passive smoke exposure [17] increases the risk of CL/P significantly. Whenever there is an effect of an environmental risk factor, it is worth exploring the potential for gene-environment interaction (GxE interaction), where the joint risk of exposure (e.g. smoking) and a genetic risk factor may be more important than the predicted marginal effects of either genes or exposures. While it is difficult to prove the existence of GxE interaction, there are some examples of possible interactions relevance to OFC. For example, the combined effects of a rare allele at TGF-alpha locus was greater than simple combinations of the marginal effects of either smoking or gene effects, suggesting GxE interaction [18, 19]. In addition, a large case/control study found women who had weekly binge drinking are at higher risk of giving birth to a child with CL/P or CP [20].

Genetic risk factors

OFCs show strong familial aggregation, which indicates a strong genetic component for this malformation. A twin study [21] from Denmark showed monozygotic (MZ) twins had higher proband-wise concordance rates for CL/P than dizygotic (DZ) twins (47% MZ twins vs. 8%

DZ twins, respectively). A similar pattern was noticed for CP. Another population-based cohort study shows higher recurrence risks with the increasingly distant degree of relatives [13]. These observations from population based twin registries suggest a high heritability for both CL/P and CP.

Linkage analysis and association analysis are two important statistical approaches to map genes for complex and heterogeneous disorders like OFC. Linkage analysis relies on multiplex families and tests for co-segregation of genetic markers (typically single nucleotide polymorphism (SNP)) and any potential gene controlling the phenotype of interest (here, CL/P or CP). Evidence of linkage to six different chromosome regions have been identified in previous linkage studies using multiplex OFC families (on chromosome 1q32, 2p13, 3q27-28, 9q21, 12p11, 14q21-24 and 16q24) [22, 23]. Particularly, SNP markers near the *IRF6* gene on chromosome 1q32 showed significant evidence of linkage for CL families and SNPs near the *FOXE1* gene on chromosome 9q21 showed significant evidence of linkage in CLP families [23].

Compared to linkage analysis, genome-wide association studies (GWAS) allow the study of millions of SNPs and consequently the identification of multiple regions throughout the genome that influence risk to OFCs (Table 1). Typically, two types of study designs are used for OFCs: the traditional case-control design and the family-based case-parent trio (triad) design. Compared to traditional case-control design, the triad design has an advantage of circumventing possible confounding due to population substructure. Whenever marker allele frequencies and baseline risk of disease vary across sub-sets of the sample (i.e. when cases and controls are drawn from genetically different sub-populations), confounding can create a biased test for association between pooled samples of cases and controls. In the triad design,

the affected children inherit alleles from their two parents and the alleles transmitted to the observed case are compared to the non-transmitted alleles in a matched case - "pseudo-control" approach. This design obviously requires the genotypes of the parents be observed, but the matched analysis means the problem of confounding is minimized because the case's alleles/genotypes are compared to alleles/genotypes in pseudo-controls possible for the given parental mating type. A transmission disequilibrium test (TDT) for alleles or for genotypes of a marker can be used to test the composite null hypothesis that there is no linkage or no linkage disequilibrium (LD) between the marker and the unobserved causal locus. This triad design may not work well for late-onset diseases since biological material from parents are needed. However, it is commonly used for childhood diseases. Studies supported by the Genes and Environment Association (GENEVA) consortium [24] represent an example of this triad design.

Until now, there have been eight GWAS for CL/P [25-32], two genome-wide meta-analysis of CL/P GWAS [7, 33] and two CP GWAS [34, 35] (Table 1). These studies have shown a high degree of genetic heterogeneity underlying risk to OFC. More than two dozen different genetic loci have been identified as influencing risk to CL/P, while only one locus has been clearly identified for CP and this association signal was limited to cases and controls of European ancestry [35]. Of these recognized genome-wide significant loci, four regions (*IRF6* on 1q32-41, and the gene desert regions on 8q24, 17q22 and 10q25.3) appear to explain 20-25% of the estimated heritability risk to CL/P [9]. However, generally for both CL/P and CP combined, all identified genetic risk regions only account for a modest proportion of the heritability of OFCs, suggesting additional genetic risk loci remain to be identified.

Table 1. A summary of genetic variants that may contribute to risk to OFC reported by previous studies.

Adapted and updated from Table 2 in Leslie and Marazita [1] and Table 1 in Beaty [9].

Locus	Candidate gene	Analysis method	Reference
1p22	***ARHGAP29***	**CL/P**	[25, 33, 36]
1p36	***PAX7***	**CL/P**	[25, 33, 37]
1p36	*MTHFR*	CL/P	[38]
10q23	*RBP4*	CL/P	[38]
1q32	***IRF6***	**CL/P**	[25, 26, 29, 30, 33, 39]
2q21	*THADA*	CL/P	[33]
2p24	*FAM49A*	CL/P	[29]
3p11	*EPHA3*	CL/P	[33]
3q12	***COL8A1/FILIPIL***	**CL/P**	[37]
3q28	*TP63*	*CL/P*	[7]
8q21	***DCAF4L2***	**CL/P**	[29, 33, 37]
8q22	*RAD54B*	CLP	[40]
8q24	***Gene desert***	**CL/P**	[25, 26, 29, 30, 39]
9q22	***FOXE1***	**CL/P, all OFCs**	[7, 23, 37]
10q25	***VAX1***	**CL/P**	[25, 29, 30]
12q13	*KRT18*	CLP	[40]
12q21	*TMEM19*	CLP	[40]
13q31	*SPRY2*	CL/P	[33]
15q22	*TPM1*	CL/P	[33]
15q24	*ARID3B*	CL/P	[29]
16p13	*CREBBP*	CL/P	[31]
17p12	***NTN1***	**CL/P**	[29, 37, 41]
17q21	*WNT9B*	CLP	[40]
17q22	*NOG*	CL/P	[29, 30, 41]
17q23	*TANC2*	CL/P	[29]
18q12	*CDH2*	CL/P	[42]
19q13	*RHPN2*	CL/P	[29]
20q12	***MAFB***	**CL/P**	[25, 29]
1p36	*GRHL3*	CP	[35, 43]
Gene x Environment			
4q22	***GRID2***	**CL/P x smoking**	[37]
9p21	***ELAVL2***	**CL/P x smoking**	[37]
8q22	***BAALC***	**CP x multivitamins**	[34]
9q31	***SMC2***	**CP x alcohol**	[34]
12q14	***TBK1***	**CP x smoking**	[34]
4p16	**SLC2A9**	**CP x smoking**	[44]
4p16	***WDR1***	**CP x smoking**	[44]

Note: findings from previous GENEVA studies are in bold-face.

GENEVA OFC Study

The GENEVA Oral Clefts project is aimed to investigate the genetic architecture of OFCs using case-parent trios collected from an international consortium. Case-parent trios were recruited by multiple investigators from Europe (Norway and Denmark), the United States (Iowa, Maryland, and Pennsylvania) and Asia (People's Republic of China, Taiwan, South Korea, Singapore, and the Philippines) [7]. Since this case-parent trio study design is robust to confounding due to population substructure, it is most appropriate to combine the trios from diverse populations into a single GWAS.

Samples were genotyped on the llumina Human610 Quadv1_B array for 589,945 SNPs, phased using SHAPEIT [45] and originally imputed with IMPUTE2 [46] software. The original genotype dataset was updated to build 37 (GRCh37) to be compatible with 1000 Genomes Phase I release (June 2011) reference panel. The original genotype data from the GENEVA case-parent trios had identified risk regions near *MAFB* and *ABCA4,* and confirmed previously identified regions such as *IRF6* and chr. 8q24 as harboring genes influencing risk for CL/P [25]. Although no genome-wide significant SNP was identified as influencing risk to CP when considered alone, there was suggestive evidence for GxE interaction controlling risk to CP [34]. For example, the study found markers in *MLLT3* increased risk of CP when the mother consumed alcohol during pregnancy [34]. A meta-analysis combining two studies, Pittsburgh Orofacial Cleft (POFC) study – which included both case-parent trios and case-control samples of OFC – and the GENEVA case-parent trio study, showed several additional risk loci including *COL8A1* (on chr. 3q12.1) may influence risk to CL/P [7].

Recently, more efficient tools have been developed for genotype imputation. The Michigan Imputation Server, which utilizes the minimac3 and the MapReduce programming model

[47], has been shown to achieve more accurate imputation while reducing computation time [47]. With increased panel size and memory requirements, it outperforms alternatives such as minimac2 [48] and IMPUTE2 [46]. Therefore, we performed genotype imputation of our GENEVA trios using the Michigan Imputation Server, and conducted a GWAS to identify the genetic risk factors for CL/P and CP.

Thesis aims

The aim of this research thesis is to identify genetic risk factors for OFCs and compare the roles of observed markers and imputed markers on identifying risk to OFCs.

Part I. Introduction to GENEVA Oral clefts database

The samples were ascertained through the GENEVA consortium which pooled case-parent trios from multiple populations into a GWAS of non-syndromic oral clefts. The aim of this consortium was to investigate the genetic variants influencing risk to OFCs, while testing for interaction between genetic markers and common environmental factors. As required by NIH, genome-wide marker data was shared through a monitored access program provided by the database for genotypes and phenotypes (dbGaP: *https://www.ncbi.nlm.nih.gov/gap*) to make these data broadly available to the scientific community. The GENEVA oral clefts study began in 2003 and was an international multi-center, case-parent trio design study. It consisted of case-parent trios collected from Europe (Norway and Denmark), the United States and Asia (People's Republic of China, Taiwan, South Korea, Singapore, and the Philippines) [7], included 1,591 CL/P complete case-parent trios (along with 318 CL/P incomplete trios) and 466 CP complete (84 CP incomplete) trios (updated by Dec, 2018). Among these trios, 668 CL/P complete (157 CL/P incomplete) trios and 215 CP complete (54 CP incomplete) trios were of European ancestry, while 895 CL/P complete (138 CL/P

incomplete) trios and 237 CP complete (22 CP incomplete) trios were of Asian ancestry (combining East Asian, Filipino and Malaysian) (Table 2). As mentioned above, DNA was genotyped on the llumina Human610 Quadv1_B array, phenotypes (e.g. type of cleft), sex, race as well as three common environmental risk factors (e.g. maternal smoking, vitamin supplementation and alcohol consumption during pregnancy) reported through direct maternal interview were used to test for GxE interaction in previous studies (Table 1). The research protocol was approved by the Institutional Review Boards (IRB) at Johns Hopkins Bloomberg School of Public Health and at each additional recruitment site. Written informed consent was obtained from both parents and assent from the case child wherever possible.

Table 2. Number of case-parent trios in the GENEVA dataset by type of OFC in the case (affected child) and ancestry of parents

	CP		CL		CLP		Total	
	Complete trios	Incomplete trios	Complete trios	Incomplete trios	Complete trios	Incomplete trios	Complete trios	Incomplete trios
European	215	54	235	53	433	104	883	211
African	1	6	0	3	3	6	4	15
Asian	231	22	214	32	675	104	1120	158
Hispanic	0	1	0	0	5	11	5	12
Native American	0	0	0	0	1	0	1	0
Malaysian	6	0	1	0	5	2	12	2
Other	13	1	7	2	12	1	32	4
All	466	84	457	90	1134	228	2057	402

Part II. Imputation on the Michigan Imputation Server

Introduction

Genotype imputation is an integral part of conducting analyses for GWAS these days. After genotyping study samples on an array usually consisting of 200,000 – 2,500,000 SNPs, imputation can expand the number of useful genetic variants by using sequencing data from a reference population (e.g. 1000 Genomes Project [49]) to identify haplotype segments shared between the observed samples and the reference panel. Genotype imputation allows efficient

inference of unobserved genotypes on a large scale [50]. It tremendously improves genome coverage by increasing SNP density, facilitates the comparison of studies originally genotyped on different SNP arrays, facilitates meta-analysis and improves the statistical power to detect associations between genetic variants and phenotypes [51]. During the past decade, imputation accuracy has been greatly improved for both common and rare alleles due to the emergence of large reference panels (e.g Haplotype Reference Consortium (HRC) [52] and 1000 Genomes Project [49]). However, this also raises concerns about computational efficiency for previous imputation tools (e.g. IMPUTE2 [46] , Beagle 4.1 [53], minimac2 [48]). For example, it would take almost one week to impute 1,000 samples using the HRC reference panel on a 100-core cluster using minimac2 [47]. To promote the computational efficiency with comparable imputation accuracy, minimac3 and minimac4 have been modified with the 'state space reduction' approach, which has yielded great improvements in imputation accuracy, run time and memory required to impute genotypes compared to previous imputation tools [47].

The Michigan Imputation Server is a cloud-based imputation server that incorporates minimac 3 [47] with a user-friendly interface. This server divides the genetic dataset into overlapping chunks and runs parallel analyses across all chunks. It performs auto-check automatically (e.g. strand orientation, file integrity, missingness and minor allele frequency distribution). If no major errors occur during the quality control process, the phased sample is imputed using one of four reference panels: HRC Version r1.1 2016 (32,470 samples and 39,635,008 sites), or Version r1 2015 (32,488 samples and 39,741,659 sites); 1000 Genomes project phase 1 (1,092 samples and 28,975,367 sites) or phase 3 (2,504 samples and 49,143,605 sites); Hapmap 2 (60 samples and 2,542,916 sites); or CAAPA-African American panel (883 samples and 31,163,897 sites), as selected by the user (submitter). This imputation

tool allows more efficient computation and yields comparable accuracy to IMPUTE2 and minimac2 when used with these large reference panels [47].

Additionally, another strategy called “pre-phasing” has been developed to ease the computational burden of imputation [54]. Pre-phasing utilizes the haplotype information specified by the family structure and transmission patterns and has been proven to yield higher efficiency when imputing a phased haplotype on two reference panels than to impute two unphased genotypes to a pair of reference haplotypes [54]. Imputation accuracy can be increased by pre-phasing with haplotyping engines such as SHAPEIT [45].

To summarize, the updated large reference panels which incorporate newly sequenced individuals and diverse variant types should increase imputation accuracy for genome-wide marker panels. This improvement is reflected in the accuracy of both imputed common and low frequency SNPs, and can facilitate analysis of rare variants. To take full advantage of these imputed marker panels with expanded genotype information, efficient imputation methods and pre-phasing are fundamental strategies to reduce computational cost and enhance studies of genetically complex diseases.

Methods

Preparing files for imputation on Michigan Imputation Server

The original genotype files were modified to pre-phase with the SHAPEIT software and then processed for imputation via the Michigan Imputation Server. The flowcharts for preparing the VCF files for imputation and individual/SNP-level filtering are summarized in Figures 1 and 2. Below is a detailed description of each step.

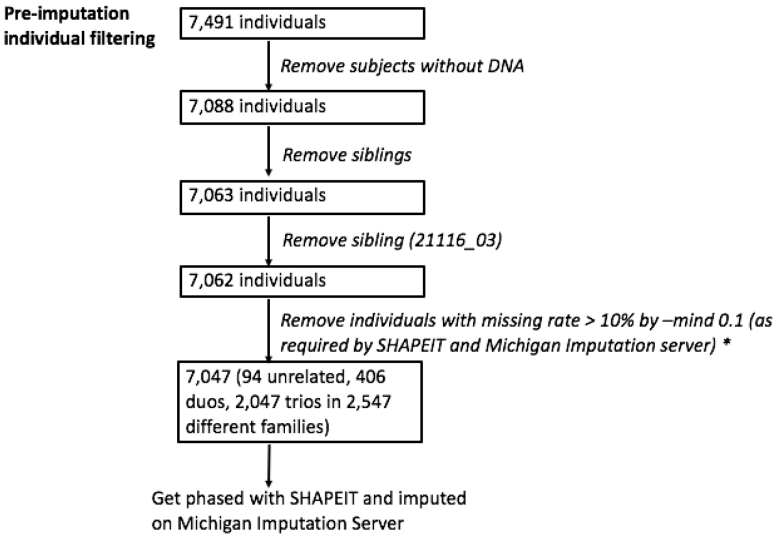


Figure 1: Flow chart for preparing files for imputation and individual level filtering. Subjects without DNA information, with SNPs missing rate >0.1 and siblings in one family were removed, resulting in 7047 individuals who were phased with SHAPEIT and imputed on Michigan Imputation Server.*

**SHAPEIT requires individuals' missing rates < 10% and Michigan Imputation Server requires all individuals have marker data for each chromosome.*

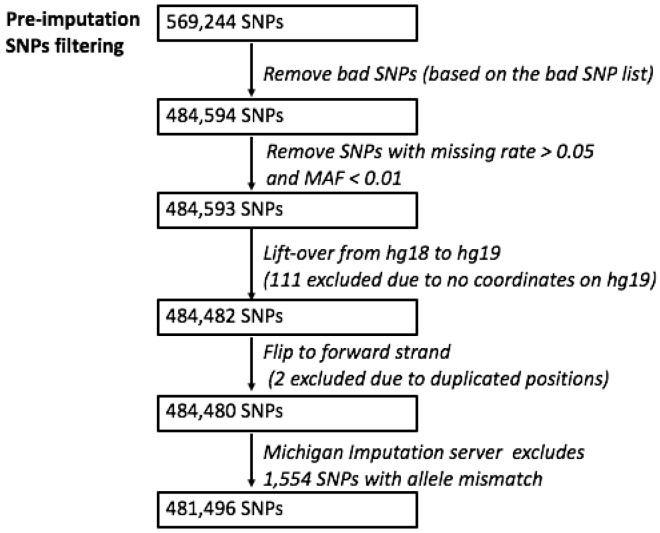


Figure 2: Flow chart for preparing the files for imputation and variant level filtering. The SNPs with high missing rates, low minor allele frequency, no coordinates on hg19 or duplicated positions were removed from this study. In addition, mismatched alleles were excluded by the Michigan Imputation Server, resulting in 481,496 SNPs being used for imputation on the Michigan Imputation Server.

Step 0. Prepare file format:

Using the original genotyped PLINK files [55] (named as "oralcleftgwas.bed", "oralcleftgwas.bim" and "oralcleftgwas.fam" under directory "jhpce01 /dcl01/beaty/data/oc_gwas/imputation_michigan/preimpute_phasing/data") from the Illumina 610Quad genotyping (which contained 569,244 SNPs on chromosomes 1-22 in 7,491 people), we obtained PLINK files in .map and .bed format for Step 1.

Step 1. Primary quality control:

The primary quality control steps were performed with PLINK 1.9 (https://www.cog-genomics.org/plink2/) [55].

Individual level QC:

Step 1.1. Remove subjects with no genotype information due to no DNA or very low quality of DNA (e.g. contamination of DNA during DNA extraction process). This step excluded 403 individuals, leaving 7,088 individuals.

Step 1.2. Remove extra siblings in each family. A total of 25 individuals were excluded, leaving 7,063 individuals.

Step 1.3. Remove a particular sibling 21116_03. Only 1 individual was excluded in this step, leaving 7062 individuals. (Note: in the "august_peds.xlsx" file, extra siblings were indicated as "4 or 5 or 6"; this sibling was indicated as "3", which is an exception.) During the original quality control procedure, individual 21116_03 was genetically found to be the sibling of 21116_01.

SNP level QC:

Step 1.4. Drop all SNPs previously flagged by original quality control (SNP missingness > 5%, Mendelian error rate > 5%, HWE< 10^{-4} and MAF < 1%). This step

resulted in rare or monomorphic SNPs being dropped from the Asian sub-group. A total of 84,650 SNPs were excluded at this step, leaving 484,594 SNPs.

Step 1.5. Drop SNPs with missing rate > 5% and minor allele frequency < 1% over all sub-groups. Only 1 SNP was excluded at this step, leaving 484,593 SNPs.

Step 2. Lift-over from hg18 to hg19 (http://github.com/sritchie73/liftOverPlink):
Original genotyped SNPs on build hg18 were "lifted over" to hg19 before imputation (currently, the Michigan Imputation Server requires hg19 coordinates). A total of 111 variants with no coordinates on hg19 were excluded, leaving 484,482 SNPs.

Step 3. Correct the information of the variant:
Correct chr4 rs100333966 minor allele to be T instead of C in the bim file. (Otherwise it would be halted on the Michigan Imputation Server.)

Step 4. Split by chromosome and flag individuals with any SNP showing a missing rate > 10%:
To phase data with SHAPEIT, individuals' missing rate for all markers must be less than or equal to 10%. Specifically, to phase the variants on chromosome 22 only, SHAPEIT required that the individual missing rate of all SNPs, which is calculated by the number of missing SNPs divided by the number of total SNPs on that chromosome for each individual, be no larger than 10%. Additionally, to impute all the chromosomes on the Michigan Imputation Server simultaneously, it is required that individuals have marker data for all chromosomes. Table 3 summarizes the list of individuals with a missing rate > 10% for each chromosome and these individuals were excluded from further analyses.

Step 5. Flip from positive strand to forward strand:

The variants were flipped to the forward strand to make them compatible with the Michigan Imputation Server. Two SNPs were excluded because they had duplicated positions, leaving 484,480 SNPs. The VCF files were converted into PLINK files with family structure information preserved.

Table 3. List of individuals with SNP missingness rate > 0.1 on individual chromosomes.

Family ID	Individual ID	Chrs with missingness rate > 10%
23042	23042_02	1, 4-9, 11, 13-20, 22
17008	17008_01	2
15096	15096_01	10-11, 16, 17, 19, 20, 22
14069	14069_02	11
23053	23053_01	16-17, 19, 22
14118	14118_01	17, 19, 22
14186	14186_01	17, 19, 22
15037	15037_01	17, 19, 21, 22
12306	12306_03	19, 22
23024	23024_02	19, 21-22
23042	23042_01	19, 21-22
23049	23049_01	19, 22
12046	12046_03	22
14114	14114_01	22
23041	23041_01	22

Note: To interpret the table, for example. individual 23042_01 had missing variant information of more than 10% on chromosomes 19, 21, and 22. Such individuals were excluded from pre-phasing and imputation.

Step 6. Remove individuals with SNPs missing rate > 10%:

We excluded 15 individuals who had SNPs missing rate > 10%, leaving 7,047 individuals.

Step 7. Phasing with SHAPEIT software:

All variant information on chromosomes 1-22 were pre-phased simultaneously using the SHAPEIT software (shapeit.v2.904.2.6.32-696.18.7.el6.x86_64) (mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html) [45]. The output files were in .haps and .sample format.

Step 8. Convert the file format to VCF files and make additional modifications:

The .haps and .sample format files generated by SHAPEIT were converted into the .vcf format. Additional modifications to these files were made as required by the Michigan

Imputation Server, including deleting the header character “chr” in the “chr” column (e.g. “chr22” became “22”).

Step 9. Check VCF files:

Check VCF files to make sure they could be imputed successfully on the Michigan Imputation Server by using script “checkVCF.py”, as recommend on the Michigan Imputation Server website.

Step 10. Sort, zip and index the files:

The files were then sorted, zipped and indexed using vcf-sort, bgzip and tabix software as required by the Michigan Imputation Server.

Step 11. Submit the prepared VCF files from step 10 to the Michigan Imputation Server:

The prepared files for chromosome 1-22 (7,047 people and 484,480 SNPs) (Table 4, Figure1, Figure 2) were imputed simultaneously on the Michigan Imputation Server. The reference panel was “1000G phase 3 v5”. Although we had trio-aware pre-phased data from SHAPEIT, Michigan Imputation Server requires choosing an option for the phasing. We chose the option “ShapeIT v2.r790” , however, it is to be noted that the server does not re-

Table 4. Comparison of the number of SNPs in the original genotyped files and the prepared files for imputation (after pre-imputation SNP filtering) for each chromosome

Chr	# of original genotyped SNPs	# of filtered genotyped SNPs for imputation
1	44,549	37,699
2	47,203	40,026
3	39,277	33,578
4	35,114	29,937
5	35,535	30,371
6	39,821	33,439
7	31,744	27,042
8	32,283	27,410
9	27,421	23,603
10	30,359	25,647
11	28,277	24,091
12	28,154	23,862
13	21,834	18,517
14	19,022	16,224
15	17,307	14,852
16	17,327	14,838
17	15,092	12,936
18	17,181	14,630
19	10,102	8,864
20	14,472	12,278
21	8,434	7,314
22	8,736	7,322
Total	569,244	484,480

Note: The total number of SNPs was reduced from 569,244 to 484,480 after the steps of pre-imputation SNP filtering. The number of SNPs remaining on each chromosome is indicated in the table above. Files containing these SNPs information were uploaded to the Michigan Imputation Server.

phase phased data. We chose the option "European" for Population (needed for internal QC purposes only) . Finally, genotype imputation was implemented by minimac3.

Quality control after imputation

The files generated from Michigan Imputation Server include .dose.vcf.gz and .info.gz for each chromosome. The flow charts for post-imputation individual and SNP-level filtering are summarized in Figures 3 and 4. Below is the description of each step.

Step 1. Explore the data and set the cut-off points for R^2:

All imputed SNPs were filtered based on $R^2 \geqslant 0.3$ with bcftools-1.9 (https://samtools.github.io/bcftools). There were 21,992,878 SNPs excluded by this cut-off value, leaving 25,108,104 SNPs. (Further exploration of different R^2 cut-off values is described in the Results section.) (Table 5)

Table 5. Comparison of the numbers of SNPs generated from imputation, filtered by R^2 and as 'hard' genotype calls.

Chr	# of SNPs post-imputation	# of SNPs after filtering R^2	# of SNPs after filtering R^2 and 'hard' call 0.1
1	3738278	1910308	1910308
2	4057648	2195643	2195643
3	3355974	1867956	1867956
4	3338298	1827487	1827487
5	3033119	1684425	1684425
6	2954483	1710246	1710246
7	2753568	1444954	1444954
8	2651635	1463683	1463683
9	2063122	1105471	1105471
10	2334121	1277810	1277810
11	2333274	1254083	1254083
12	2242777	1206916	1206916
13	1661713	951726	951726
14	1525694	824510	824510
15	1404183	704418	704418
16	1549341	724282	724282
17	1345848	607045	607045
18	1319664	722217	722217
19	1084557	453904	453904
20	1047637	544673	544673
21	653809	321435	321435
22	652239	304912	304912
Total	47100982	25108104	25108104

Note: There were substantial SNPs excluded by setting R^2 threshold at 0.3. No SNPs excluded by hard call threshold of 0.1.

Step 2. Make the 'hard' genotype call in PLINK software:

The VCF format was changed into PLINK format and 'hard' genotype calls were made by setting threshold

0.1 within the PLINK 2 (https://www.cog-genomics.org/plink/2.0/) software. If the calls have uncertainty greater than 0.1 (genotype likelihoods smaller than 0.9), they were treated as missing; and the rest were regarded as hard calls [55]. No additional SNPs were excluded from this step, therefore leaving 25,108,104 SNPs for analysis. (Table 5, Figure 3)

Step 3. Modify the format in the resulting PLINK files to retain the family structure: Family information is lost when PLINK files are converted into VCF files. To convert VCF files back to PLINK files, the family, individual, maternal and paternal IDs were modified with reference to the original genotyped "oralcleftgwas.fam" file to retain family information needed for downstream analysis.

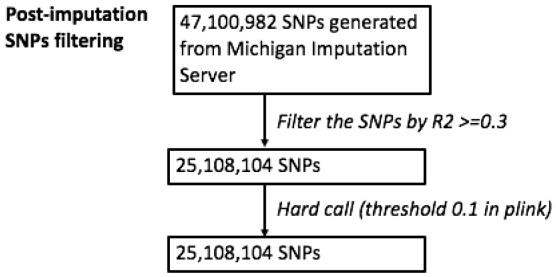


Figure 3: Flow chart for post-imputation variant level filtering. The imputed SNPs were further filtered by R2 and 'hard' genotype call.

Step 4. Quality control exploration (Results shown in the Results section):

Step 4.1. Check for individual missingness rates and SNP missingness rates on the entire cohort.

Step 4.2. Split by ethnicity (East Asian+Malaysian+Filippino vs. European)

Step 4.2.1. Check for SNP missingness rate

Step 4.2.2. Check for MAF and HWE only in founders

Step 5. Remove individuals duplicated in the POFC dataset:

A total of 393 individuals were removed due to overlap with POFC dataset, leaving 6,654 individuals. This step was implemented so that GENEVA study may be meta-analyzed with POFC study in subsequent GWAS analyses.

Step 6. Keep only complete trios:

Eight hundred and seven individuals were removed due to the incomplete trio status as indicated in the "august_peds.xlsx" list, leaving 5,847 individuals in 1,949 complete trios.

Step 7. Remove any incomplete trio created by pre-imputation "Step 6. Remove individuals with SNPs missing rate > 10%" :

A total of 21 individuals were removed due to the incomplete trios produced by excluding individuals with SNPs missing rate > 10%, leaving 5,826 individuals in 1,942 trios.

Step 8. Double check to make sure the duplicated siblings had been removed and check for Mendelian errors.

Step 9. Split individuals by OFC phenotype in the affected child (CL/P & CP) and by ethnicity (Asian & European):

After splitting the population by OFC phenotype and racial group, there were 235 Asian complete CP trios (705 individuals), 891 Asian complete CL/P trios (2,673 individuals), 203 complete European CP trios (609 individuals) and 575 complete European CL/P trios (1,725 individuals) (Figure 4).

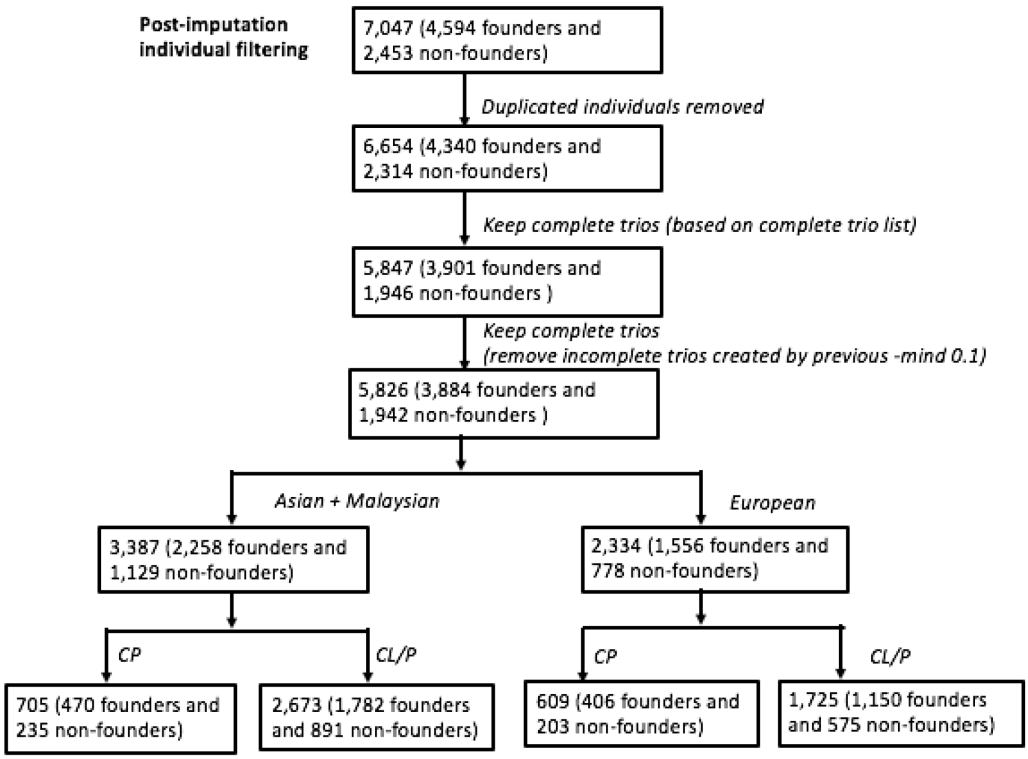


Figure 4: Flow chart for post-imputation individual level filtering. Individuals duplicated across the Geneva and POFC datasets, and incomplete trios were removed from the study. The number of individuals split by phenotype and racial group are summarized.

Results

1. Explore the R^2 to determine the cutoff value:

R^2 is the estimated correlation between the imputed and true genotyped variants, and can be regarded as an indicator of imputation accuracy. As indicated from Table 6, a substantial number of SNPs had a R^2 of either below 0.1 or above 0.9. R^2 below 0.1 suggests very poor imputation quality. The occurrence of poor imputation often correlates with the SNP's minor allele frequency. From Figure 5, we can see rare variants with MAF < 1% are more likely to have poor imputation quality. With reference to the previous GWAS study, which used a

Table 6. SNPs remaining by R^2 cutoff

R^2 cutoff	SNPs remaining
0.0	47,099,551
0.1	32,517,090
0.2	28,571,598
0.3	25,108,104
0.4	21,828,637
0.5	18,731,933
0.6	15,818,178
0.7	13,059,060
0.8	10,400,849
0.9	7,601,183
1.0	15,465

cutoff of 0.5 [7] and the observed distribution of R^2 in our GENEVA data, we decided to choose a cut-off point of 0.3, which left 25,108,104 SNPs available for analysis.

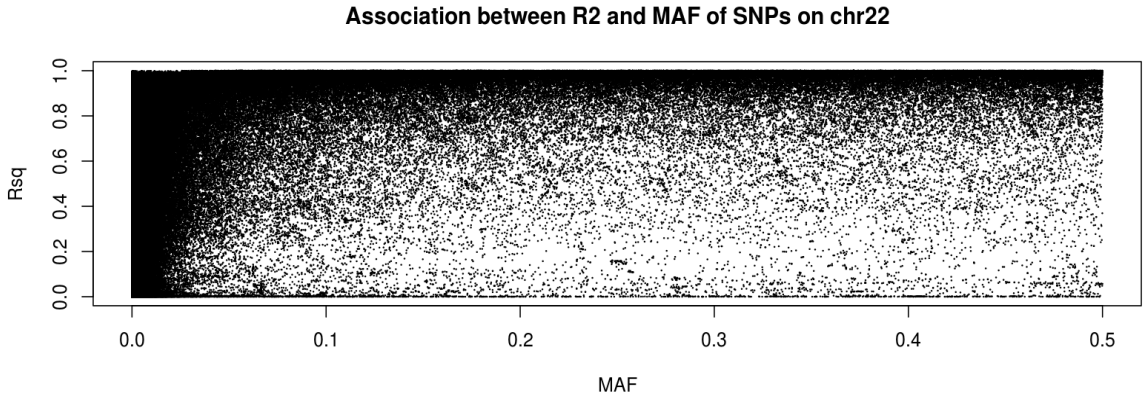


Figure 5: The association between imputation accuracy (R2) and minor allele frequency (MAF) for all SNPs on chromosome 22. The SNPs with lower R2 values tended to be rare variants (MAF $\leqslant$ 0.05).

2. Explore the MAF distribution for all SNPs on chromosome 22 in Asian and European groups separately after filtering by R^2:

Considering the large amount of data and computational burden, we explored the MAF distribution among Asian and European ancestry groups separately on chromosome 22 after filtering all imputed SNPs by R^2. Among the 304,912 SNPs on chromosome 22, there were 196,464 rare variants with MAF < 1% and 24,660 variants with MAF between 1-5%, 83,788 common variants with MAF $\geqslant$5% in the Asian sub-group, while there were 177,733 rare variants with MAF < 1% and 33,450 variants with MAF between 1-5%, and 93,729 common variants with MAF $\geqslant$ 5% in the European sub-group. The MAF distribution in these two racial groups (Asian and European) were generally similar (Figure 6).

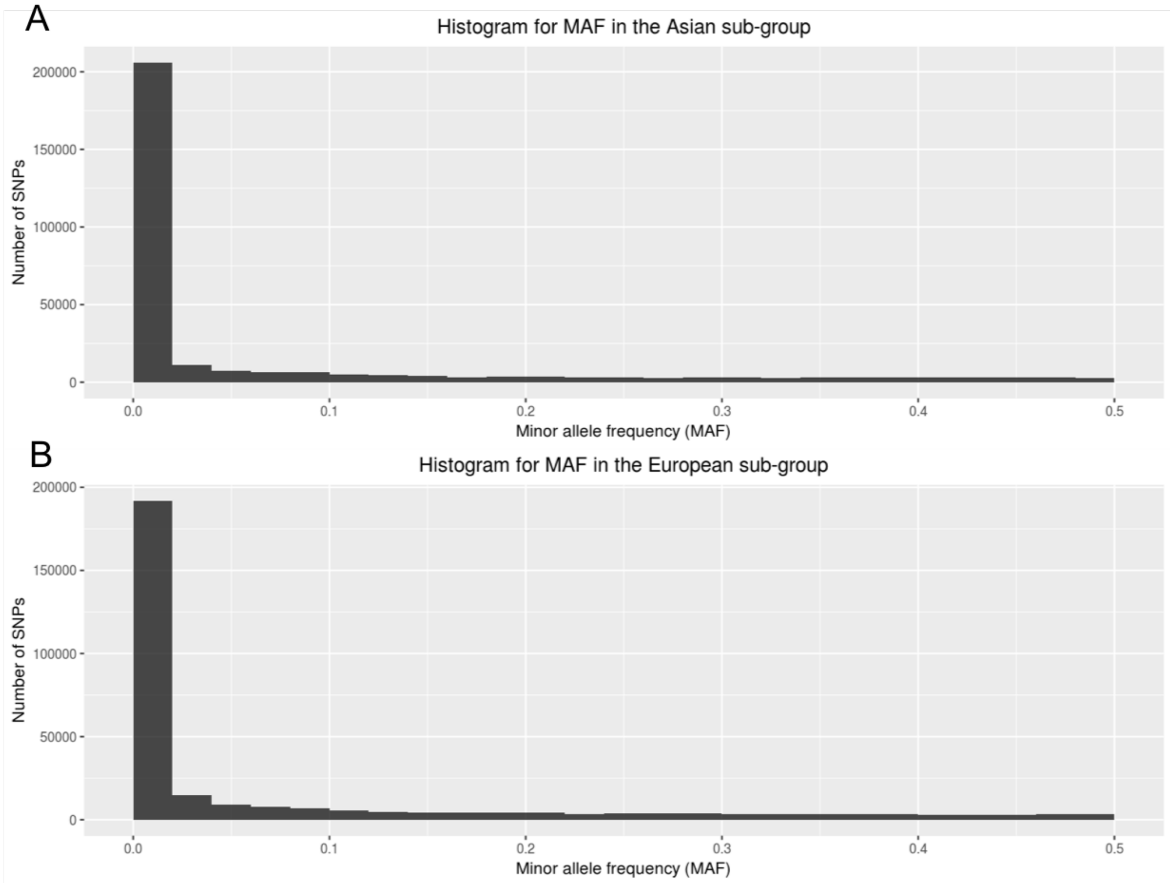


Figure 6: The distribution of minor allele frequency in the Asian (A) and European (B) subgroups. Similar distributions of MAF was noticed between the two sub-groups.

3. Exploring deviation from HWE for all SNPs on chromosome 22 in Asian and European parents separately:

Either 100 or 55 SNPs would have been excluded from the 304,912 imputed SNPs on chromosome 22 by setting the cut-off point for deviation from HWE 10^{-4} or 10^{-5}, respectively, in the Asian sub-group.

Table 7. Number of SNPs remaining by p-value of Hardy-Weinberg equilibrium cutoff values.

	Asian	European
P-value of test for HWE	# of SNPs	# of SNPs
$\geqslant 10^{-2}$	302,582	302,744
$\geqslant 10^{-3}$	304,475	304,614
$\geqslant 10^{-4}$	304,812	304,868
$\geqslant 10^{-5}$	304,857	304,897
$\geqslant 10^{-6}$	304,912	304,909
$\geqslant 10^{-7}$	304,912	304,912

There were no SNPs violating HWE at a p-value point of 10^{-6} in the Asian sub-group. In comparison with the European sub-group, 44 or 15 SNPs would have been excluded from the 304,912 SNPs on chromosome 22 by setting the cut-off point of HWE 10^{-4} or 10^{-5} in the

European sub-group, respectively. There were no SNPs violating HWE at cut-off point of 10^{-7} in the European group (Table 7). We set the HWE threshold at 5 x 10^{-7} (i.e., SNPs with HWE p-value < 5 x 10^{-7} are excluded).

4. Explore the missingness and Mendelian error rate after all the pre-imputation and post-imputation quality control steps.

Both the missingness rate and Mendelian errors decreased to 0 after pre-phasing with SHAPEIT. This may be due to the fact that SHAPEIT detects Mendelian errors and set these genotype values to missing during phasing, yielding zero Mendelian errors. Due to pre-phasing with family structure information with SHAPEIT, all Mendelian errors were set to missing. Missingness remained low (missingness for both SNPs and individuals are 0) as SNPs were filtered before imputation.

Discussion

The genome-wide marker data was pre-phased and imputed using the Michigan Imputation Server. Essential QC steps were performed both pre-imputation and post-imputation to ensure the high quality of the resulting dataset of over 25 million markers. The final dataset contains 25,108,104 SNPs (including both common and rare variants) for 5,826 complete case-parent trios (including 3,387 Asian and 2,334 European individuals) which were then prepared for downstream analysis.

Part III. Genome-Wide Association analysis

Introduction

Case-parent trio design

As an alternative to population-based (e.g. case-control) studies considering independent individuals, family-based (e.g. case-parent trio) designs are frequently used to detect the association between genetic variants and disease-related phenotypes. In the case-parent trio design, both the affected child (case) and their parents are genotyped, and "pseudo-controls" are created by the alleles or genotypes that could have been transmitted from the observed parental mating type. The allelic TDT becomes a matched comparison of alleles transmitted to the case versus those not transmitted, and McNemars' chi-square test can be used to test the null hypothesis of no deviation from strict Mendelian transmission to the affected child [56]. In the genotypic TDT, the observed genotype of the case is compared to the other three "pseudo-control" genotypes that could have been generated by the parental mating under a 1:3 matched design [57].

There are some unique advantages of this case-parent trio design. First, it allows tests for any parent-of-origin effect (e.g. imprinting effect) by comparing phenotypic effect of maternal allele vs. the paternal allele [58]. Second, it is easier to identify *de novo* variants using this case-parent trio design. Third, population stratification is circumvented by drawing pseudo-controls from the observed parental mating type. In contrast, spurious associations may arise due to population stratification in case-control studies, whenever the allele frequency of markers differs across sub-populations of cases and controls. Last, case-parent trio design can be more powerful for rare diseases compared with case-control studies [59].

However, there are some drawbacks to this study design. Since the case-parent trio design requires the recruitment of both parents and children, early-onset diseases are better candidates for this design, as it is more feasible to collect parents' genetic information in early-onset compared to late-onset diseases. Additionally, it is not possible to test for independent environmental factors contributing to risk of disease because the cases and the pseudo-controls share any maternal exposure status as they come from the same family.

To summarize, the case-parent trio design has some advantages for investigating OFC, a common birth defect but a relatively rare disorder with a complex and heterogeneous etiology. Birth defects are not common occurrences and parents are generally available to provide DNA. Additionally, because this study design is robust to population stratification, it is a good design for investigating the etiology of OFC where case-parent trios are recruited from multiple distinct populations (e.g. Asians and Europeans).

Genotypic transmission disequilibrium test analysis

The transmission disequilibrium test (TDT) is a fundamental approach for testing genetic associations of a disease phenotype under the case-parent trio design [56]. The null hypothesis of the TDT is a composite of no association and no linkage between the observed markers and an unobserved causal gene. This test focuses on the departure from Mendelian expectations for the marker in a sample strictly ascertained through an affected child and checks if a target allele is preferentially transmitted to the affected child from a heterozygous parent. For the allelic TDT, McNemar's chi-square test is used to compare the number of times the target allele is transmitted to the affected child with the number of times the alternative allele is transmitted to the affected child [58]. This test is a non-parametric method with no assumption of the distribution of the phenotype (e.g. if the phenotype follows normal

distribution in the study population) nor the specific model of inheritance, and is also referred to as “allelic TDT” [56].

Another frequently used alternative approach to the allelic TDT is the genotypic transmission disequilibrium test (gTDT). In the gTDT analysis, the genotype of the affected individual is compared to the three pseudo-control genotypes that could occur given the parents’ genotypes [60]. Conditional logistic regression can be used to account for this matching structure [61]. Compared to the allelic TDT analysis, the gTDT analysis treats the trio as a family unit, assuming a specific genetic model (i.e. either an additive, recessive or dominant mode of inheritance). Closed-form solutions have been developed to estimate the coefficient of the conditional logistic regression models under different genetic architectures [57]. Moreover, because the underlying genetic model is usually unknown, the maximum over all gTDT statistics can be used as a test statistic to evaluate the effect of any one SNP. A fast approach to compute this test statistic as well as a permutation-based p value has been proposed [57, 62]. In addition, the gTDT allows the calculation of an odds ratio of the effect of the variant on the outcome and an associated confidence interval, which facilitates the combination of results from multiple trio studies, while the allelic TDT only provides chi-square test statistic values and their p-values. The gTDT also allows tests for GxE interaction [57] and has been proven to be more powerful than the allelic TDT analysis in some circumstances [63], which makes gTDT attractive for case-parent trio studies. The gTDT procedures can be implemented using the open source Bioconductor *trio* package, available at https://bioconductor.org/packages/release/bioc/html/trio.html. Therefore, we analyzed the case-parent trios from the GENEVA consortium using gTDT (as implemented in the *trio* package) to test for linkage and association with all SNPs (observed and imputed) passing QC steps discussed above to identify genes influencing risk to OFC.

Multiple comparisons

Since a large number of SNPs (observed and imputed) are used in this study, we must consider the issue of multiple comparisons. Bonferroni correction is the most common approach for tackling the multiple comparisons issue. This method calculates the corrected type I error rate by dividing the traditional threshold for declaring statistical significance (0.05) by the number of independent comparisons. However, this method is conservative if the traditional threshold is divided by the number of SNPs in the GWAS analysis because the SNPs are correlated and the number of independent comparisons is way smaller than the total number of SNPs [64]. In this study, we used a threshold of 5×10^{-7} to declare genome wide significance [29].

Methods

The gTDT analyses for CP and CL/P trios separately were performed using the *trio* package (version 3.20.0) on common SNPs (i.e. those with MAF $\geqslant$ 5%) in the combined set of all trios together and then stratified into Asian and European sub-groups. Manhattan plots and QQ plots were created for each analysis to show signals from the gTDT and to check for potential bias in the test statistic. The signals from the combined imputed and genotyped SNPs were compared with signals from the original observed genotype SNPs only. We expect genotype imputation to increase the sensitivity for detecting and defining chromosomal regions showing evidence of harboring genes controlling risk to OFCs.

SNPs reaching genome wide significance from the gTDT analysis were annotated with an online tool SNPnexus (https://snp-nexus.org) [65] to indicate potentially important genes. LocusZoom plots (http://locuszoom.org/) [66] were created to show evidence of linkage and association for the genome-wide significant regions indicated by the gTDT analysis under an

additive model in the combined set of all trios, plus the European and Asian sub-groups separately, with reference to genome build hg19 and 1000 Genomes European/Asian populations (Nov 2014). For the LocusZoom LD information, because the choice of European/Asian reference populations did not make much difference in the observed LD patterns in either the European or Asian sub-group of trios, we reported regions of interest from the gTDT analysis in the combined set of all trios using Europeans as a reference population, while the plots of European sub-group used the European population as reference, and plots of Asian sub-group used the Asian population as reference.

Results

The gTDT analysis on autosomal SNPs in 1,942 CL/P trios from all the populations showed locations of potential causal polymorphisms (Figure 7A, Supplementary Figure 1A). The Q-Q plot also showed significant evidence of association and linkage beyond what can be explained by random chance alone (Figure 8). The gTDT analysis of CP trios did not yield any signal of association (Supplementary Figure 2). This is not unexpected given the small sample size of CP trios, which limits the statistical power to detect CP associated genetic variants. Previous GWAS of GENEVA study had failed to map genes influencing risk to CP [34] while a meta-analysis of GENEVA and POFC studies identified only one gene significantly associated with CP [7]. On the other hand, 639 SNPs (47 genotyped, 592 imputed) from nine different regions showed genome-wide significance ($p < 5 \times 10^{-7}$) for CL/P. Among them, eight nearby genes have been previously reported, including 8q24 (gene desert region), 1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*). One locus (18q12, nearest gene *TTR*) which has not been detected previously also reached genome-wide significance ($p = 4.33 \times 10^{-8}$). We stratified the gTDT analysis by ethnic group to check for consistency of these significantly associated loci.

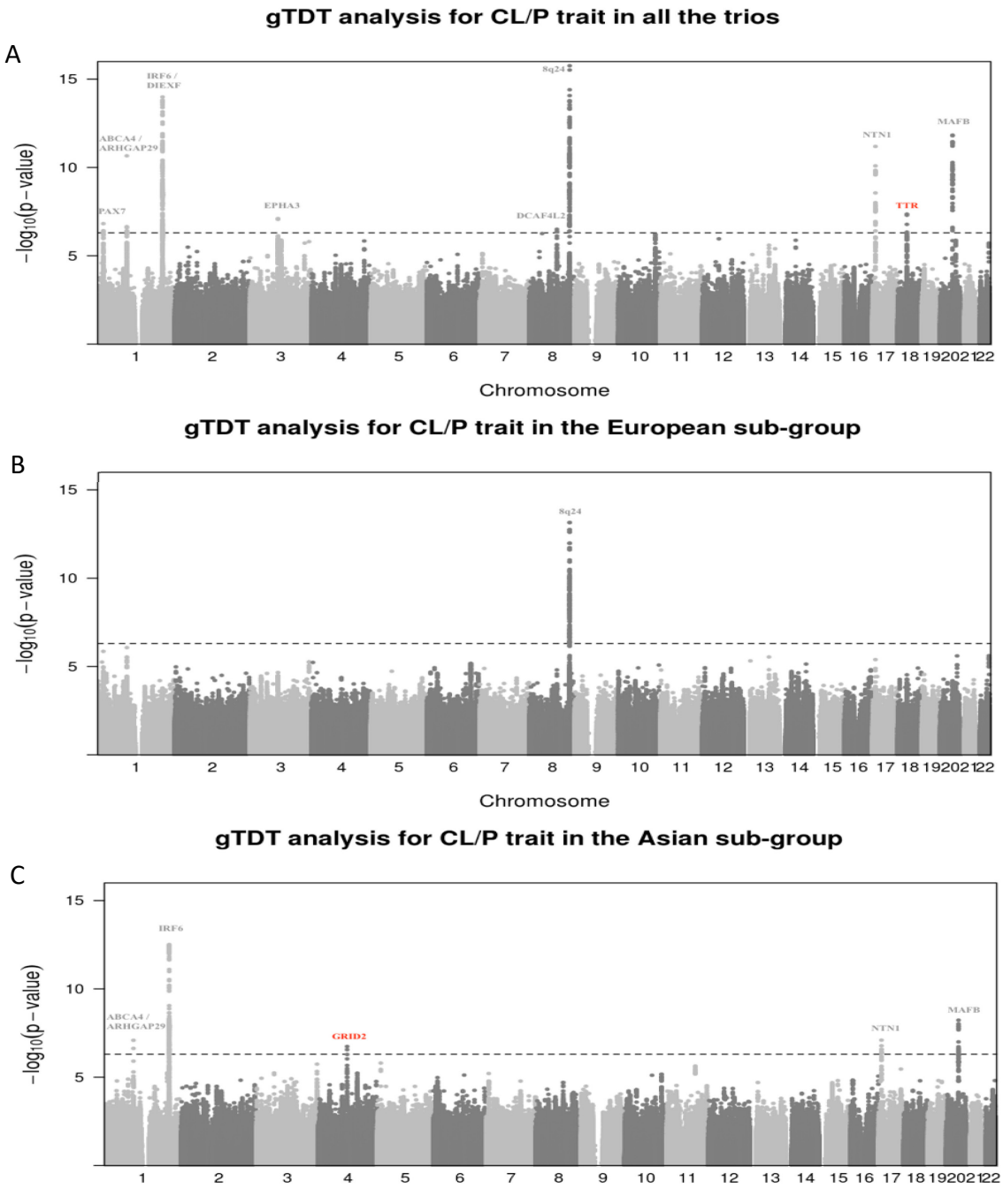


Figure 7: Manhattan plot for gTDT analysis (imputed + genotyped SNPs) of CL/P trait in the combined set of all trios (A), European (B) and Asian (C) sub-groups. Peaks are labeled with overlapped genes or closest upstream or downstream gene in the region. Grey labels indicate the previously reported loci, and red labels indicate new loci.

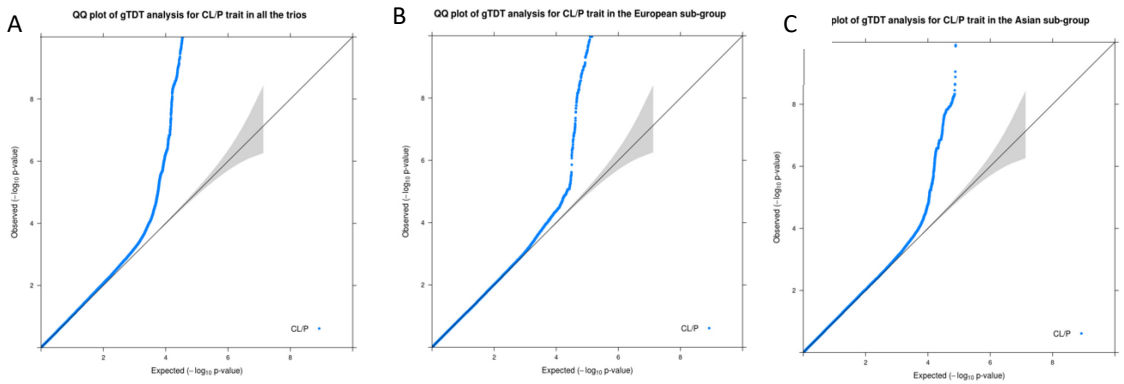


Figure 8: Q-Q plot of all autosomal SNPs (imputed + genotyped SNPs) in the combined set of all trios (A), European (B) and Asian (C) sub-groups. X axis is the negative logarithm of the expected p value whereas the y axis is the negative logarithm of the observed p value. The gray shaded region indicates 95% confidence interval. A departure from the expected value was observed after taking random chance into consideration.

Region 8q24 showed genome-wide evidence only in the 575 trios of European ancestry, while 1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*) were significant only in the larger group of 891 Asian ancestry trios (Figures 7B, C). Additionally, a novel locus (4q22, *GRID2*) yielded genome-wide significance ($p = 1.82 \times 10^{-7}$) in the trios of Asian ancestry alone (Figure 7C). Numerous SNPs in the 8q24 region showed genome-wide significance in this gTDT analysis. The most significant SNP was rs17242358 in the combined set of all trios ($p = 1.75 \times 10^{-16}$). SNP rs17242358 showed over-transmission of the A allele (over the G allele) with estimated relative risk (RR) = 2.09 [95%CI: (1.76, 2.49)] (Table 8). This estimated RR is similar in both ethnic groups: European trios gave an estimated RR = 2.09 [95%CI: (1.72, 2.54)] and Asian trios gave a similar estimated RR = 2.14 [95%CI: (1.38, 3.32)]. However, this locus reached genome-wide significance in the European sub-group only ($p = 7.11 \times 10^{-14}$) while achieving nominal significance in the Asian sub-group ($p = 0.00073$). The LD and p-value patterns of SNPs around rs17242358 are similar between the combined set of all CL/P trios (when using 1000Genomes of European ancestry as the reference population) and European sub-group (using Europeans as the reference population), and quite distinct from Asian sub-group (using Asians as the reference population) (Figure 9). The MAF for rs17242358 was 23% in European parents and 2% in

Asian parents, a substantial difference which would severely limit statistical power to detect any signal in the latter sub-group.

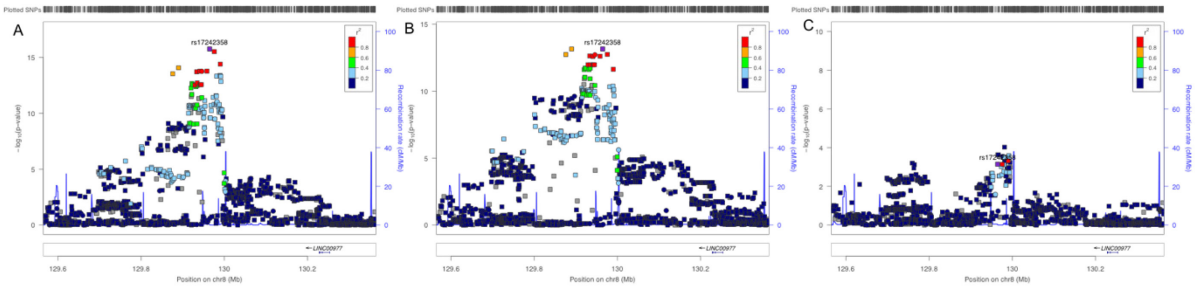


Figure 9: LocusZoom plot for GENEVA CL/P gTDT analysis results. The peak SNP (8q24.21, rs17242358) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 8 is labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European ancestry trios only. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian ancestry only. Linkage disequilibrium was color coded with reference to Asian populations.

In contrast to the 8q24 region, which showed consistent signals in the combined set of all trios and the European sub-group, the regions 1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*) were genome-wide significant in the combined set of all trios and in the Asian sub-group. For example, the lead SNP in the 1q32 (*IRF6*) region in the combined set of all trios was rs12075674. The RR of CL/P when comparing trios with A allele at this SNP to those without was 0.57 [95%CI: (0.50, 0.66)] as calculated in the combined set of all trios (Table 8). This RR is similar to that estimated in the Asian trios only with RR = 0.58 [95%CI: (0.51, 0.68)] and European trios with RR = 0.39 [95%CI: (0.20, 0.75)]. However, this SNP reached genome-wide significance only in the combined set of all trios ($p = 1.02 \times 10^{-14}$) and in the Asian sub-group ($p = 3.30 \times 10^{-13}$). The European trios showed nominal significance with p-value of 5.2×10^{-3}. The p-value and LD patterns are consistent between the combined set of all CL/P trios (with LD reference to European populations) and Asian sub-group (with LD reference to Asian populations), rather than European sub-group (with LD reference to European populations) (Figure 10). The MAFs of this allele in the Asian sub-group and European sub-group are 36% and 2% respectively, meaning the statistical power to detect any linkage and LD would be much lower in the European sub-group. Similarly, the

most-significant SNPs at the 20q12 (*MAFB*) and 17p13 (*NTN1*) loci, which were rs6072084 and rs12944377, showed consistent p-values and LD patterns in the combined set of all trios and the Asian trios (Figures 11, 12). The lead SNPs at each of the above four regions were all imputed SNPs except for 1p22 (*ABCA4*), which was directly genotyped. The p-values and LD pattern in the 1p22 (*ABCA4*) region was distinctive between the combined set of all trios and the Asian trios considered alone (Figure 13) despite the most significant SNP (rs560426) reaching genome-wide significance in both groups.

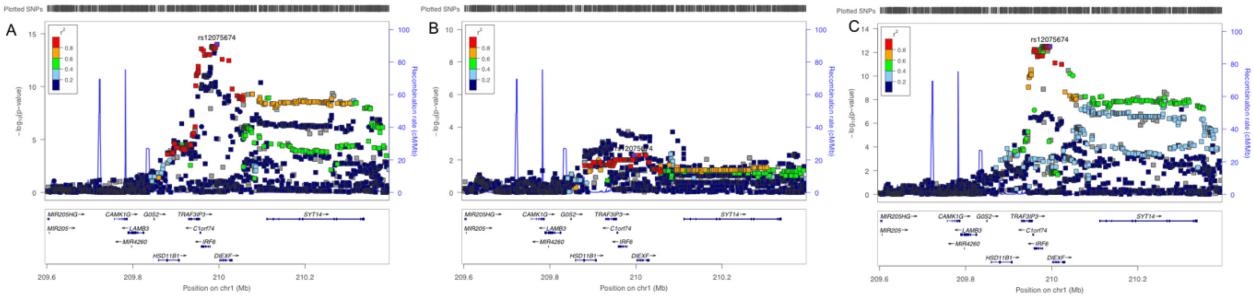


Figure 10: LocusZoom plot for GENEVA CL/P gTDT analysis results. One of the lead SNPs (1q32.2, rs12075674) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 1 was labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

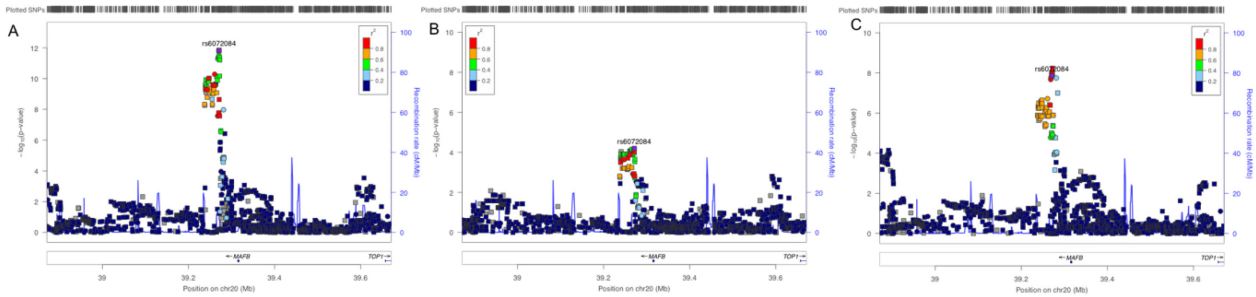


Figure 11: LocusZoom plot for GENEVA CL/P gTDT analysis results. The lead SNP (20q12, rs6072084) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 20 was labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

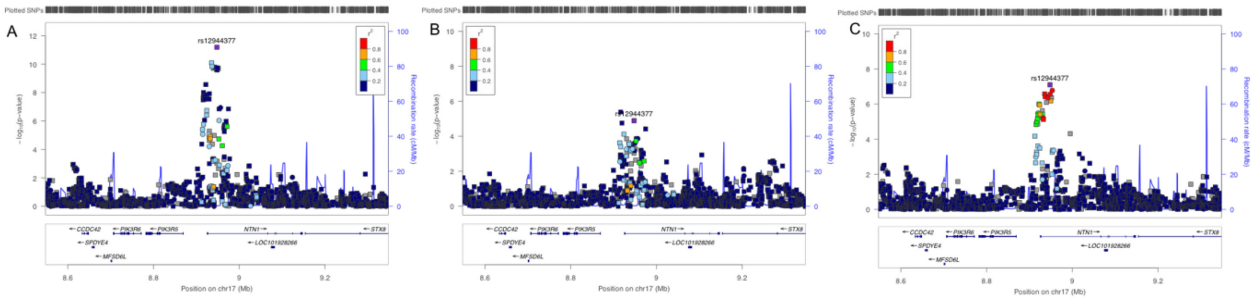


Figure 12: LocusZoom plot for GENEVA CL/P gTDT analysis results. The lead SNP (17p13.1, rs12944377) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 17 was labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

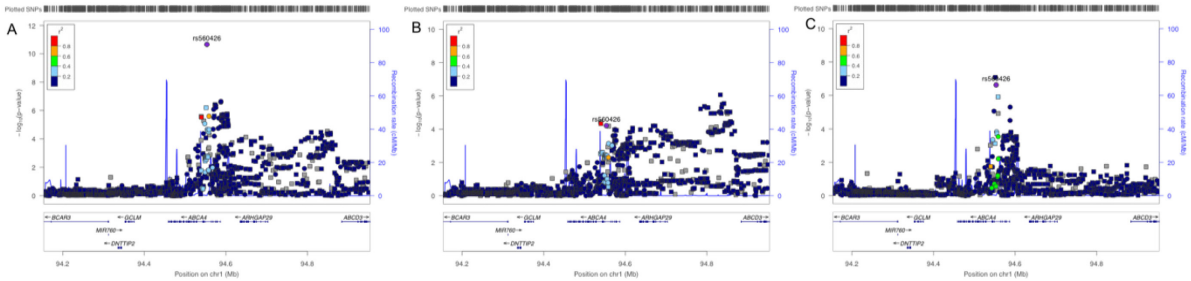


Figure 13: LocusZoom plot for GENEVA CL/P gTDT analysis results. One of the most significant SNPs (1p22.1, rs560426) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 1 was labeled. (A). gTDT analysis in all the populations. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

SNPs near 18q12 (*TTR*) and at 4q22 (*GRID2*) reached genome-wide significance in the combined set of all trios and in the Asian trios considered separately. A 32Kb region around the *TTR* gene encompassing 58 SNPs contained 4 imputed SNPs yielding genome-wide significance, among which 3 SNPs were located 15Kb – 10Kb upstream of TTR and 1 SNP was located 17Kb downstream of this gene. The most significant SNP in this region on 18q12 (*TTR*) was rs1375445. The estimated RR for CL/P based on the combined set of all trios contributed by rs1375445 was 1.35 [95%CI: (1.21, 1.51)] ($p = 4.33 \times 10^{-8}$) (Table 8). No genome-wide significance was detected in the European ($p = 2.94 \times 10^{-5}$) and Asian trios ($p = 5.52 \times 10^{-5}$) when considered separately. Patterns of p-value and LD around rs1375445 were different in the combined set of all trios, the European trios and the Asian trios (Figure 14). Additionally, 2,798 SNPs overlapped with *GRID2* gene and 2 imputed SNPs reached

genome-wide significance. The most significant SNP in this region on 4q22 (*GRID2*) was rs1471079. The estimated RR for CL/P in the Asian trios based on genotype at rs1471079 was 0.70 [95%CI: (0.60, 0.80)] ($p = 1.82 \times 10^{-7}$) (Table 8). No genome-wide significance was detected in the combined set of all trios ($p = 4.99 \times 10^{-5}$) or in the European trios ($p = 0.93$) alone. The p-value and LD patterns were similar between the combined set of all trios and Asian trios, but quite different in the European trios (Figure 15). Both of these SNPs (on *TTR* and *GRID2* genes) were imputed with high accuracy (rs1375445: $R^2 = 0.96$; rs1471079: $R^2 = 0.97$).

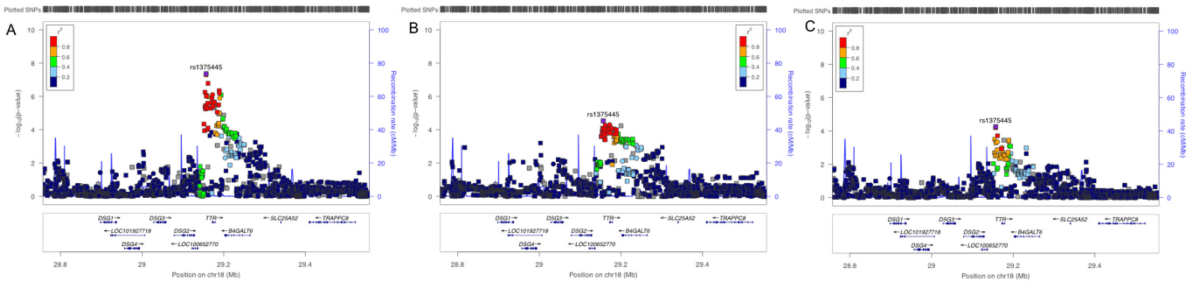


Figure 14: LocusZoom plot for GENEVA CL/P gTDT analysis results. The lead SNP (18q12.1, rs1375445) (as calculated by gTDT analysis in the combined set of all trios) on chromosome 18 was labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

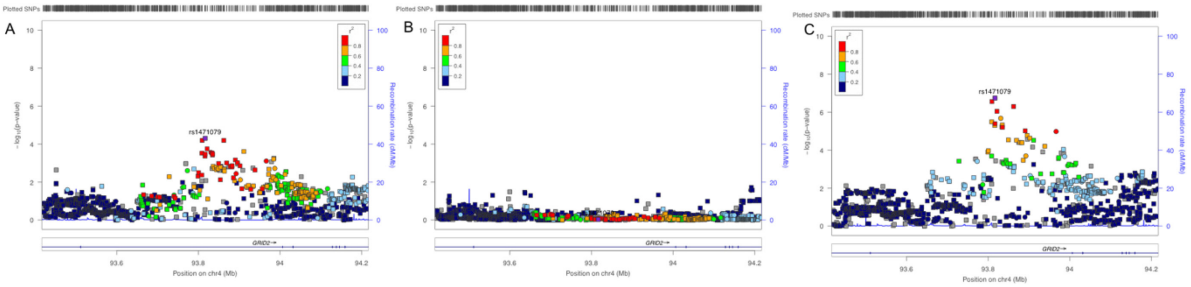


Figure 15: LocusZoom plot for GENEVA CL/P gTDT analysis results. The lead SNP (4q22.2, rs1471079) (as calculated by gTDT analysis in the Asian populations) on chromosome 4 was labeled. (A). gTDT analysis in the combined set of all trios. Linkage disequilibrium was color coded with reference to European populations. (B). gTDT analysis in European populations. Linkage disequilibrium was color coded with reference to European populations. (C). gTDT analysis in Asian populations. Linkage disequilibrium was color coded with reference to Asian populations.

Overall, the ability to detect potentially causal genes was greatly increased by genotype imputation (Figure 7, Supplementary Figure 1). Several loci (e.g. in 1p36.13, *PAX7*) reached genome-wide significance when the gTDT analysis used both genotyped and imputed SNPs

compared to analysis of genotyped SNPs alone (Figure 7 and Supplementary Figure 1). Moreover, 12 of the 15 lead SNPs in each of these regions were imputed (Table 8). Statistical significance was increased for each locus by using both genotyped and imputed SNPs for GWAS analysis. For instance, the most significant p-value for markers in 17p13.1 (*NTN1*) from the gTDT analysis combining imputed and genotyped SNPs is 6.46×10^{-12} (lead SNP: rs12944377, imputed) while it is 2.07×10^{-8} (lead SNP: rs9788972, genotyped) for gTDT analysis with genotyped SNPs only.

Table 8. Top significant SNPs from gTDT analysis of CL/P in the combined set of all trios, in the European and the Asian subgroups.

	Chr	Position	SNP Name	Locus	Gene	p-value	Genotyped (R^2)	RR	SE	Minor Allele	Major Allele	MAF	European MAF	Asian MAF
CL/P_ALL	1	18986508	rs56075776	1p36.13	*PAX7*	1.52E-07	Imp (0.93)	1.49	0.08	A	G	0.20	0.42	0.04
	1	94553438	rs560426	1p22.1	*ABCA4/ ARHGAP29*	2.20E-11	Gen	1.44	0.05	C	T	0.39	0.47	0.33
	1	209995470	rs12075674	1q32.2	*DIEXF*	1.02E-14	Imp (0.99)	0.57	0.07	A	G	0.22	0.02	0.36
	3	89534377	rs7632427	3p11.1	*EPHA3*	7.76E-08	Gen	0.71	0.06	C	T	0.27	0.38	0.18
	8	88868340	rs12543318	8q21.3	*DCAF4L2*	3.20E-07	Gen	0.76	0.05	A	C	0.48	0.38	0.38
	8	129964873	rs17242358	8q24.21	*Gene desert*	1.75E-16	Imp (0.97)	2.09	0.09	A	G	0.11	0.23	0.02
	17	8947708	rs12944377	17p13.1	*NTN1*	6.46E-12	Imp (0.98)	1.50	0.06	T	C	0.37	0.42	0.22
	18	29156999	rs1375445	18q12.1	*TTR*	4.33E-08	Imp (0.96)	1.35	0.06	T	C	0.36	0.36	0.36
	20	39271400	rs6072084	20q12	*MAFB*	1.46E-12	Imp (0.98)	0.68	0.05	A	C	0.47	0.45	0.42
CL/P_ EUROPEAN	8	129890188	rs17241253	8q24.21	*Gene desert*	7.09E-14	Imp (0.96)	2.18	0.10	C	T	0.09	0.21	0.01
CL/P_ ASIAN	1	94551450	rs17461953	1p22.1	*ABCA4/ ARHGAP29*	8.09E-08	Imp (0.64)	2.04	0.13	C	A	0.08	0.08	0.08
	1	209978777	rs17015250	1q32.2	*IRF6*	3.06E-13	Imp (0.99)	0.60	0.07	G	T	0.45	0.36	0.48
	4	93816799	rs1471079	4q22.2	*GRID2*	1.82E-07	Imp (0.97)	0.70	0.07	A	C	0.44	0.45	0.43
	17	8947708	rs12944377	17p13.1	*NTN1*	7.92E-08	Imp (0.98)	1.54	0.08	T	C	0.37	0.42	0.22
	20	39272739	rs4812449	20q12	*MAFB*	5.84E-09	Imp (0.97)	0.66	0.07	G	C	0.41	0.39	0.43

Note: Imp = Imputed; Gen = Genotyped; MAF = Minor allele frequency in the combine set of all trios; European MAF = Minor allele frequency in the European sub-group; Asian MAF = Minor allele frequency in the Asian sub-group.

Discussion

We identified two novel regions yielding genome-wide significant evidence for CL/P [18q12 (*TTR*), 4q22 (*GRID2*)] and replicated previous findings for multiple genes [e.g. 8q24, 1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*)] [7, 25, 29]. The stratified analysis of European and Asian sub-groups also recapitulated the prior findings in that the signal from markers on 8q24 was most significant in European populations while the other loci [1q32 (*IRF6*), 20q12 (*MAFB*), 17p13 (*NTN1*) and 1p22 (*ABCA4*)] were more significant in Asian populations [7, 9, 25, 29]. Stratification by sub-group also revealed a new locus on 4q22 (*GRID2*) achieving genome-wide significance in the Asian sub-group. These findings likely reflect differences in MAFs between these two sub-populations [67]. For example, the MAF for rs17242358 (8q24) was 23% in European trios and 2% in Asian trios, resulting in fewer informative Asian trios, which would decrease the statistical power to detect significant associations in Asian trios only. Moreover, most of the lead SNPs of these significant regions were imputed, suggesting increased statistical power to detect potential risk loci was achieved through genotype imputation using more recent larger reference panel.

Despite the fact that the most significant SNP in the 8q24 region (rs17242358) did not reach genome-wide significance in the 891 trios of Asian ancestry (where the MAF was low), the estimated RR for CL/P was still high [RR = 2.14, 95%CI: (1.38, 3.32)], and yielded nominal significance (p = 0.00073). This RR value was higher when compared to the previous GWAS study using allelic TDT analysis with 1,038 trios of Asian ancestry, which showed OR(case) = 1.42 [95%CI: (1.08, 1.85)] under additive model and p-value of transmission = 8.9×10^{-3} [25]. For trios of European ancestry, the variants in this locus showed similar effect sizes in our study compared to the previous study [25]. This suggests a greater detection power of the gTDT analysis compared to the allelic TDT analysis [63], especially for rare variants in a particular population. Additionally,

the Michigan Imputation Server allows more efficient use of large reference panels, which also increases the efficiency and accuracy of imputation compared to previous imputation tools [47].

TTR is expressed in liver and pancreas. It encodes transthyretin, a homo-tetrameric carrier protein, which is responsible for transporting thyroid hormones and retinol in the plasma [68]. Clinical syndromes related to *TTR* are familial amyloid polyneuropathy (FAP) and cardiomyopathy (http://omim.org/entry/176300). Despite the absence of clinical reports about CL/P being associated with *TTR*, disorders resulting from thyroid hormones have been detected in CL/P patients [69], which may suggest some indirect role on risk to CL/P modulated by thyroid hormones levels.

GRID2 has previously shown evidence of gene and smoking interactions in the European trios, but the gene itself has never shown genome-wide significance [37]. However, in our study, the lead SNP of this locus reached genome-wide significance in the Asian trios ($p = 1.82 \times 10^{-7}$) but not in European trios ($p = 0.93$), although the MAF is similar between Asians and Europeans with value of 43% and 45% respectively. This may result from different LD patterns between sub-groups. It may also be due to the prevalence of maternal smoking or maternal exposure to smoking in the two sub-groups. Further investigation is needed to understand if *GRID2* harbors a causal locus for CL/P. The protein encoded by *GRID2* is a group of ionotropic glutamate receptors, expressed at cerebellar Purkinje cells, ovary and testis (https://www.ncbi.nlm.nih.gov/gene/2895). Mutations in this gene can cause cerebellar ataxia. No direct evidence has been detected for association between markers in this gene and risk to CL/P, but it may affect the reproductive cells by interrupting the function of their membrane receptors.

Our findings confirm the complex and genetic heterogeneity of OFCs. Using 25 million imputed and observed common SNPs, we detected previously reported genes yielding evidence of linkage and association in this case-parent trio study. We also identified two new loci achieving genome-wide significance which will require further investigation. Stratification by ethnic sub-groups (Asian and European ancestry) helped to detect risk loci controlling risk to OFCs in specific sub-groups while imputation helped increase the power to identify such risk loci.

Limitations

We detected two novel sites from the gTDT analysis in this case-parent trio study. However, this genome-wide association study using the case-parent trio design cannot directly indicate if the SNP is casual, in LD with unobserved casual locus or a spurious signal due to some confounding factors. Therefore, further replication and functional studies are needed to confirm these findings. In addition, it is still challenging to interpret the biological functions of these significant SNPs. Moreover, we are unable to detect rare variants which may be associated with OFC malformations using this GWAS approach. Finally, since we included only biallelic variants for analysis, we were unable to detect the association between other types of variants and risk of OFCs.

Public health impact

These findings deepen our understanding of the complex genetic architecture of OFCs, which may differ across populations. They help further investigate the potential pathways associated with risk of OFC.

Acknowledgment

This work was partially funded by NIH/NIDCR grant R03DE027121. JHPCE computing resources supported by the Department of Biostatistics were used to carry out this study.

I sincerely thank all the families participating in this international study. This work wouldn't have been possible without their dedication.

I am especially indebted to my advisors Dr. Terri H. Beaty and Dr. Debashree Ray, who have been incredibly supportive of my thesis work and my career goals. The door to their office was always open whenever I ran into a trouble spot. I wish to thank them for their patience helping me to structure and revise my thesis. They have taught me more than academics and shown me what a good scientist (and person) should be.

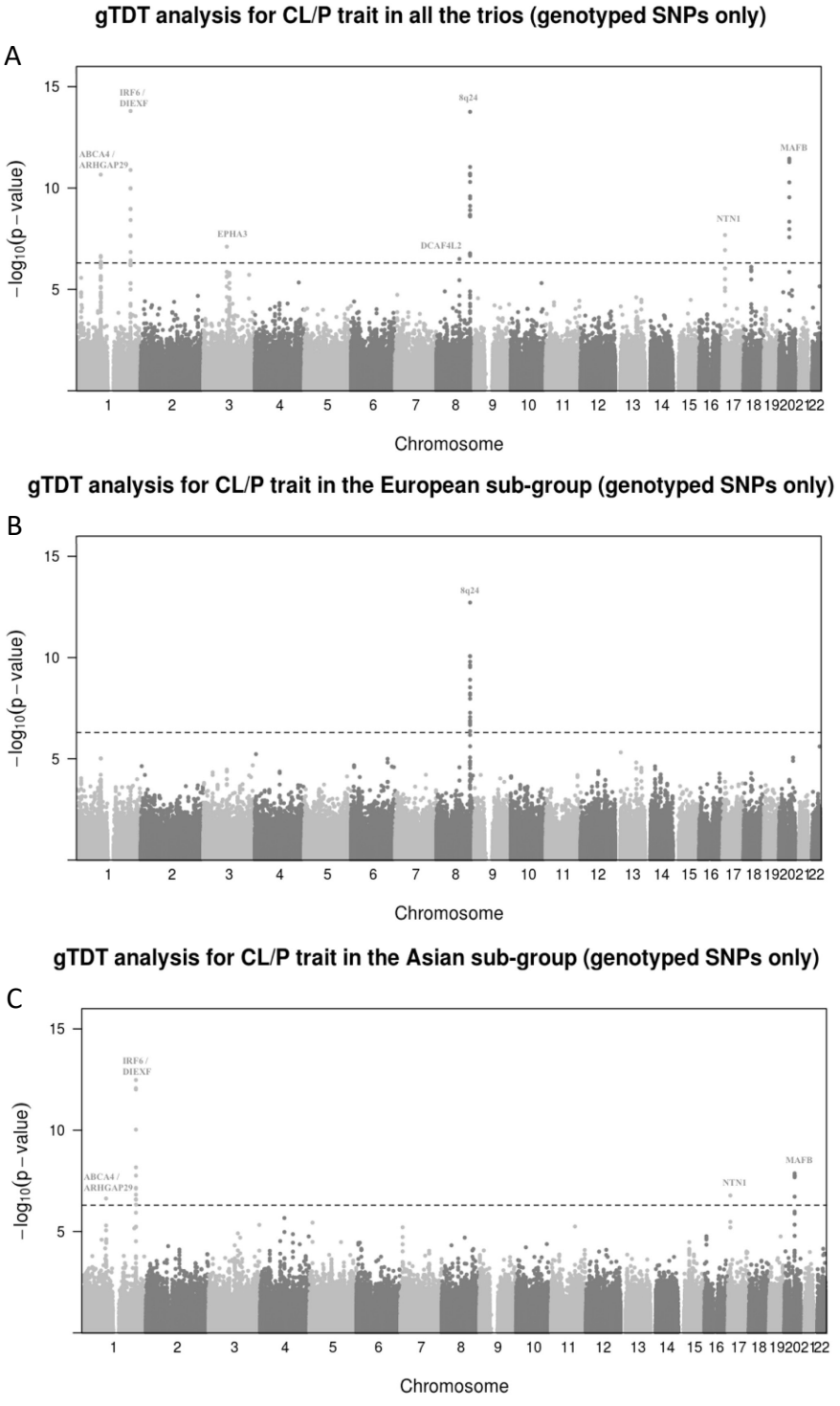
I would also like to thank my thesis reader: Dr. Margaret A. Taub, for her thorough reading of this thesis and for her insightful comments and encouragement.

I must also express my profound gratitude to all of those whom I had the pleasure to work with and who have been assisting me in this thesis work and other projects. I would especially like to thank Ms. Jacqueline A. Bidinger, Dr. Margaret Taub, Dr. Candelaria Coggiano, Sowmya Venkataraghavan, Cristian Valencia and Yuangen Liu, who have provided me extensive professional guidance on this work. Without their precious support and assistance, this thesis work would not have been conducted smoothly.

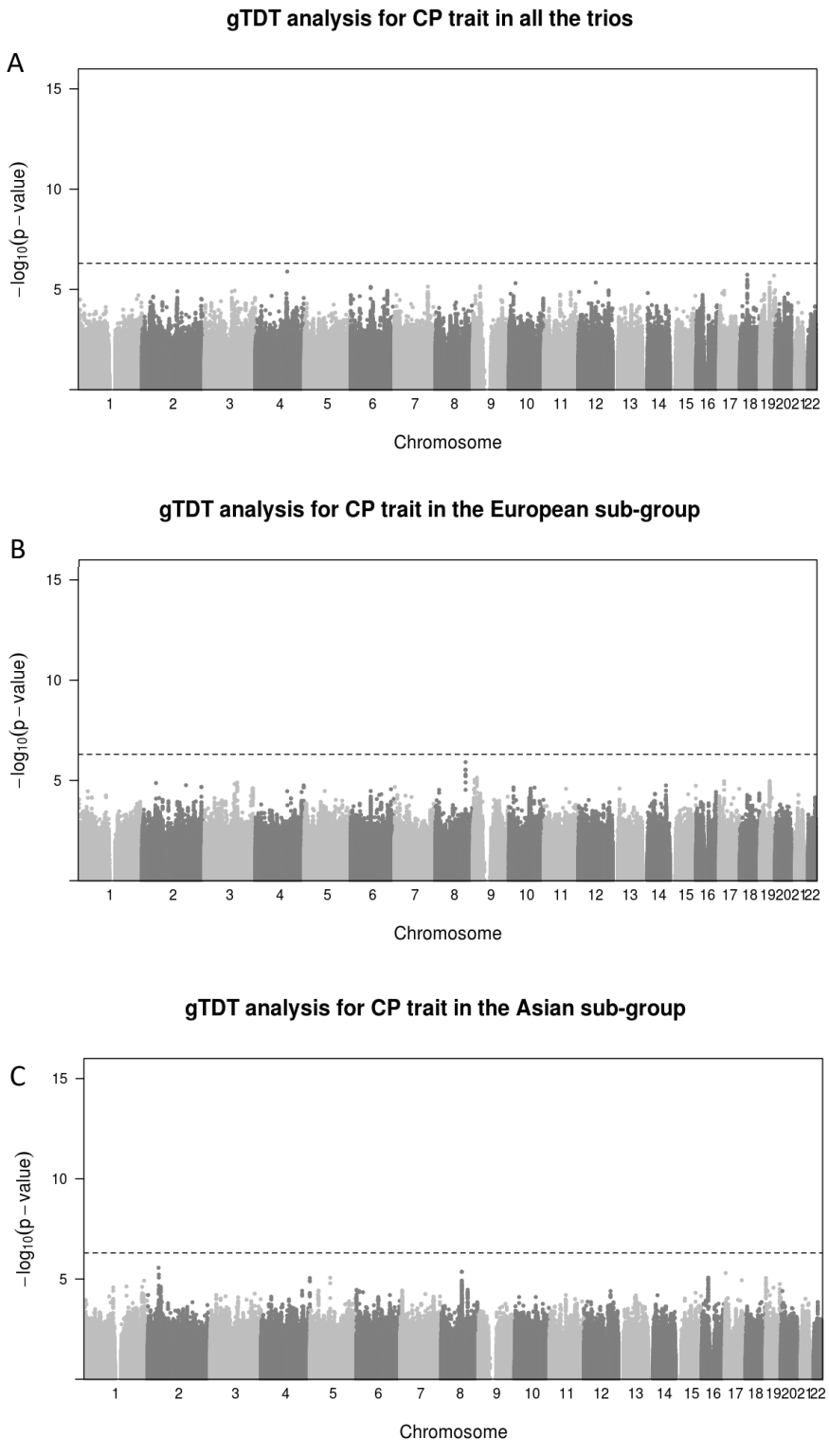
My sincere thanks also go to my classmates, fellows and mentors from genetic epidemiology track. I feel very motivated surrounded by these diligent, dedicated and supportive people and learned a great deal from them.

Finally, I extend my deepest thank to my families: my father, mother, younger brother and my husband. It was my parents' unconditional love, care and support that allowed me to pursue study abroad. I am profoundly grateful to my husband, who has been standing by me and providing unending inspiration throughout my study at JHSPH and writing this thesis, even though 7,000 miles are between us.

Appendix



Supplementary Figure 1: Manhattan plot for gTDT analysis (genotyped SNPs alone) of CL/P trait in the combined set of all trios (A), European (B) and Asian (C) sub-groups. Peaks are labeled with overlapped genes or closest upstream or downstream gene in the region. Grey labels indicate the previously reported loci. The horizontal dash line indicates the genome-wide significant p value of 5×10^{-7}.



Supplementary Figure 2: Manhattan plot for gTDT analysis (imputed and genotyped SNPs) of CP trait in the combined set of all trios (A), European (B) and Asian (C) sub-groups. Peaks are labeled with overlapped genes or closest upstream or downstream gene in the region. The horizontal dash line indicates the genome-wide significant p value of 5×10^{-7}.

References

1. Leslie, E.J. and M.L. Marazita, *Genetics of Orofacial Cleft Birth Defects.* Curr Genet Med Rep, 2015. **3**(3): p. 118-126.
2. Yu, W., M. Serrano, S.S. Miguel, et al., *Cleft lip and palate genetics and application in early embryological development.* Indian J Plast Surg, 2009. **42 Suppl**: p. S35-50.
3. Shi, M., G.L. Wehby, and J.C. Murray, *Review on genetic variants and maternal smoking in the etiology of oral clefts and other birth defects.* Birth Defects Res C Embryo Today, 2008. **84**(1): p. 16-29.
4. Young, D.L., R.A. Schneider, D. Hu, et al., *Genetic and teratogenic approaches to craniofacial development.* Crit Rev Oral Biol Med, 2000. **11**(3): p. 304-17.
5. Suazo, J., J.L. Santos, A. Colombo, et al., *Gene-gene interaction for nonsyndromic cleft lip with or without cleft palate in Chilean case-parent trios.* Arch Oral Biol, 2018. **91**: p. 91-95.
6. Estandia-Ortega, B., J.A. Velazquez-Aragon, M.A. Alcantara-Ortigoza, et al., *5,10-Methylenetetrahydrofolate reductase single nucleotide polymorphisms and gene-environment interaction analysis in non-syndromic cleft lip/palate.* Eur J Oral Sci, 2014. **122**(2): p. 109-13.
7. Leslie, E.J., J.C. Carlson, J.R. Shaffer, 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): p. 275-286.
8. Christensen, K., K. Juel, A.M. Herskind, et al., *Long term follow up study of survival associated with cleft lip and palate at birth.* BMJ, 2004. **328**(7453): p. 1405.
9. Beaty, T.H., M.L. Marazita, and E.J. Leslie, *Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities.* F1000Res, 2016. **5**: p. 2800.
10. Mai, C.T., C.H. Cassell, R.E. Meyer, et al., *Birth defects data from population-based birth defects surveillance programs in the United States, 2007 to 2011: highlighting orofacial clefts.* Birth Defects Res A Clin Mol Teratol, 2014. **100**(11): p. 895-904.
11. Mossey, P.A. and B. Modell, *Epidemiology of oral clefts 2012: an international perspective.* Front Oral Biol, 2012. **16**: p. 1-18.
12. IPDTOC working group, *Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC).* Cleft Palate Craniofac J, 2011. **48**(1): p. 66-81.

13. Grosen, D., C. Chevrier, A. Skytthe, et al., *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, 2010. **47**(3): p. 162-8.
14. Martinez-Frias, M.L., *Topiramate in pregnancy: preliminary experience from the UK Epilepsy and Pregnancy Register.* Neurology, 2009. **72**(23): p. 2054-5; author reply 2055.
15. Moody, M., O. Le, M. Rickert, et al., *Folic acid supplementation increases survival and modulates high risk HPV-induced phenotypes in oral squamous cell carcinoma cells and correlates with p53 mRNA transcriptional down-regulation.* Cancer Cell Int, 2012. **12**: p. 10.
16. Honein, M.A., O. Devine, S.D. Grosse, et al., *Prevention of orofacial clefts caused by smoking: implications of the Surgeon General's report.* Birth Defects Res A Clin Mol Teratol, 2014. **100**(11): p. 822-5.
17. Kummet, C.M., L.M. Moreno, A.J. Wilcox, et al., *Passive Smoke Exposure as a Risk Factor for Oral Clefts-A Large International Population-Based Study.* Am J Epidemiol, 2016. **183**(9): p. 834-41.
18. Hwang, S.J., T.H. Beaty, S.R. Panny, et al., *Association study of transforming growth factor alpha (TGF alpha) TaqI polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects.* Am J Epidemiol, 1995. **141**(7): p. 629-36.
19. Shaw, G.M., C.R. Wasserman, E.J. Lammer, et al., *Orofacial clefts, parental cigarette smoking, and transforming growth factor-alpha gene variants.* Am J Hum Genet, 1996. **58**(3): p. 551-61.
20. Romitti, P.A., L. Sun, M.A. Honein, et al., *Maternal periconceptional alcohol consumption and risk of orofacial clefts.* Am J Epidemiol, 2007. **166**(7): p. 775-85.
21. Grosen, D., C. Bille, I. Petersen, et al., *Risk of oral clefts in twins.* Epidemiology, 2011. **22**(3): p. 313-9.
22. Marazita, M.L., J.C. Murray, A.C. Lidral, et al., *Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35.* Am J Hum Genet, 2004. **75**(2): p. 161-73.
23. Marazita, M.L., A.C. Lidral, J.C. Murray, 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): p. 151-70.

24. Cornelis, M.C., A. Agrawal, J.W. Cole, et al., *The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions.* Genet Epidemiol, 2010. **34**(4): p. 364-72.
25. Beaty, T.H., J.C. Murray, M.L. Marazita, 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): p. 525-9.
26. Birnbaum, S., K.U. Ludwig, H. Reutter, et al., *Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24.* Nat Genet, 2009. **41**(4): p. 473-7.
27. Camargo, M., D. Rivera, L. Moreno, et al., *GWAS reveals new recessive loci associated with non-syndromic facial clefting.* Eur J Med Genet, 2012. **55**(10): p. 510-4.
28. Grant, S.F., K. Wang, H. Zhang, et al., *A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24.* J Pediatr, 2009. **155**(6): p. 909-13.
29. Leslie, E.J., J.C. Carlson, J.R. Shaffer, 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): p. 2862-2872.
30. Mangold, E., K.U. Ludwig, S. Birnbaum, et al., *Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate.* Nat Genet, 2010. **42**(1): p. 24-6.
31. Sun, Y., Y. Huang, A. Yin, et al., *Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate.* Nat Commun, 2015. **6**: p. 6414.
32. Wolf, Z.T., H.A. Brand, J.R. Shaffer, et al., *Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate.* PLoS Genet, 2015. **11**(3): p. e1005059.
33. Ludwig, K.U., E. Mangold, S. Herms, et al., *Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci.* Nat Genet, 2012. **44**(9): p. 968-71.
34. Beaty, T.H., I. Ruczinski, J.C. Murray, et al., *Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate.* Genet Epidemiol, 2011. **35**(6): p. 469-78.

35. Leslie, E.J., H. Liu, J.C. Carlson, et al., *A Genome-wide Association Study of Nonsyndromic Cleft Palate Identifies an Etiologic Missense Variant in GRHL3.* Am J Hum Genet, 2016. **98**(4): p. 744-54.
36. Paul, B.J., K. Palmer, J.C. Sharp, et al., *ARHGAP29 Mutation Is Associated with Abnormal Oral Epithelial Adhesions.* J Dent Res, 2017. **96**(11): p. 1298-1305.
37. Beaty, T.H., M.A. Taub, A.F. Scott, et al., *Confirming genes influencing risk to cleft lip with/without cleft palate in a case-parent trio study.* Hum Genet, 2013. **132**(7): p. 771-81.
38. Zhang, S.J., P. Meng, J. Zhang, et al., *Machine Learning Models for Genetic Risk Assessment of Infants with Non-syndromic Orofacial Cleft.* Genomics Proteomics Bioinformatics, 2018. **16**(5): p. 354-364.
39. Carlson, J.C., M.A. Taub, E. Feingold, et al., *Identifying Genetic Sources of Phenotypic Heterogeneity in Orofacial Clefts by Targeted Sequencing.* Birth Defects Res, 2017. **109**(13): p. 1030-1038.
40. Yu, Y., X. Zuo, M. He, et al., *Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity.* Nat Commun, 2017. **8**: p. 14364.
41. Leslie, E.J., M.A. Taub, H. Liu, 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): p. 397-411.
42. Kumari, P., S.K. Singh, and R. Raman, *A novel non-coding RNA within an intron of CDH2 and association of its SNP with non-syndromic cleft lip and palate.* Gene, 2018. **658**: p. 123-128.
43. Eshete, M.A., H. Liu, M. Li, et al., *Loss-of-Function GRHL3 Variants Detected in African Patients with Isolated Cleft Palate.* J Dent Res, 2018. **97**(1): p. 41-48.
44. Wu, T., H. Schwender, I. Ruczinski, et al., *Evidence of gene-environment interaction for two genes on chromosome 4 and environmental tobacco smoke in controlling the risk of nonsyndromic cleft palate.* PLoS One, 2014. **9**(2): p. e88088.
45. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2011. **9**(2): p. 179-81.
46. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.

47. Das, S., L. Forer, S. Schonherr, et al., *Next-generation genotype imputation service and methods.* Nat Genet, 2016. **48**(10): p. 1284-1287.
48. Fuchsberger, C., G.R. Abecasis, and D.A. Hinds, *minimac2: faster genotype imputation.* Bioinformatics, 2015. **31**(5): p. 782-4.
49. Genomes Project, C., A. Auton, L.D. Brooks, et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
50. Porcu, E., S. Sanna, C. Fuchsberger, et al., *Genotype imputation in genome-wide association studies.* Curr Protoc Hum Genet, 2013. **Chapter 1**: p. Unit 1 25.
51. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
52. McCarthy, S., S. Das, W. Kretzschmar, et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
53. Browning, B.L. and S.R. Browning, *Genotype Imputation with Millions of Reference Samples.* Am J Hum Genet, 2016. **98**(1): p. 116-26.
54. Howie, B., C. Fuchsberger, M. Stephens, et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.
55. Chang, C.C., C.C. Chow, L.C. Tellier, et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
56. Spielman, R.S., R.E. McGinnis, and W.J. Ewens, *Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).* Am J Hum Genet, 1993. **52**(3): p. 506-16.
57. Schwender, H., M.A. Taub, T.H. Beaty, et al., *Rapid testing of SNPs and gene-environment interactions in case-parent trio data based on exact analytic parameter estimation.* Biometrics, 2012. **68**(3): p. 766-73.
58. Cordell, H.J., B.J. Barratt, and D.G. Clayton, *Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects.* Genet Epidemiol, 2004. **26**(3): p. 167-85.
59. Laird, N.M. and C. Lange, *Family-based designs in the age of large-scale gene-association studies.* Nat Rev Genet, 2006. **7**(5): p. 385-94.
60. Laird, N.M. and C. Lange, *Association analysis in family designs*, in *The Fundamentals of Modern Statistical Genetics.* 2011, Springer New York: New York. p. 139-170.

61. Self, S.G., G. Longton, K.J. Kopecky, et al., *On estimating HLA/disease association with application to a study of aplastic anemia.* Biometrics, 1991. **47**(1): p. 53-61.
62. Schwender, H., Q. Li, C. Neumann, et al., *Detecting disease variants in case-parent trio studies using the bioconductor software package trio.* Genet Epidemiol, 2014. **38**(6): p. 516-22.
63. Schaid, D.J., *Likelihoods and TDT for the case-parents design.* Genet Epidemiol, 1999. **16**(3): p. 250-60.
64. Goeman, J.J. and A. Solari, *Multiple hypothesis testing in genomics.* Stat Med, 2014. **33**(11): p. 1946-78.
65. Dayem Ullah, A.Z., N.R. Lemoine, and C. Chelala, *SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update).* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W65-70.
66. Pruim, R.J., R.P. Welch, S. Sanna, et al., *LocusZoom: regional visualization of genome-wide association scan results.* Bioinformatics, 2010. **26**(18): p. 2336-7.
67. Murray, T., M.A. Taub, I. Ruczinski, et al., *Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between Asians and Europeans.* Genet Epidemiol, 2012. **36**(4): p. 392-9.
68. Saraiva, M.J., *Transthyretin mutations in hyperthyroxinemia and amyloid diseases.* Hum Mutat, 2001. **17**(6): p. 493-503.
69. Akin, M.A., S. Kurtoglu, D. Sarici, et al., *Endocrine abnormalities of patients with cleft lip and/or cleft palate during the neonatal period.* Turk J Med Sci, 2014. **44**(4): p. 696-702.

Wanying Zhang, MD

3501 Saint Paul Street Apt 543, Baltimore, MD
410-528-3168 • wzhang93@jhmi.edu

EDUCATION

Master of Science 2017-2019
Johns Hopkins Bloomberg School of Public Health
Department of Epidemiology

Doctor of Medicine 2009-2017
Peking Union Medical College

Bachelor of Science 2009-2012
Tsinghua University
Department of Life Sciences

AWARDS AND HONORS

- **Master's Tuition Scholarship** 2018 – 2019
  Awarded to students in good academic standing pursuing the second year of study in a two-year master's program by Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health.
- **National Scholarship** 2012 – 2016
  Awarded annually to the overall excellent student (1-3 out of 83 students in Class 2009 of Peking Union Medical College) by the National Ministry of Education
- **Pacemaker to Merit Student & Model Student of Academic Records** 2014 – 2015
  Awarded annually to the academic excellent student (top1-7 out of 83 students in Class 2009 of Peking Union Medical College) by the National Ministry of Education
- **Awards of Excellence in National College Student Innovation Plan** 2011 – 2012
  Awarded annually to the most innovative research project by the Education Committee of Peking Union Medical College. (I was awarded for applying a method of combining sequencing with high-resolution melting to detect pathogenic gene mutation in Primary Hypertrophic Osteoarthropathy patients.)

TRAINING

Intern 05/2014 – 11/2016
Peking Union Medical College Hospital
Clinical rotations in Internal Medicine, Surgery, Gynecology, Pediatrics, Radiology, Anaethesiology, Neurology.

Exchange Student 04/2016 – 05/2016
University of Michigan Hospital
Clinical rotations in Anaethesiology

Intern 8/2015 – 09/2015
Cleveland Clinic
Observership in Pain management

RESEARCH EXPERIENCE

Research Assistant 06/2018 - present
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
- Genetic imputation on Michigan Imputation Server
- Genome-wide association study on orofacial cleft disease based on GENEVA cohort data

Research Assistant 03/2016 – 11/2016
Department of Human Anatomy, Histology and Embryology, Peking Union Medical College
- National Alzheimer Disease Epidemiology Study (published)

Statistical Analyst 08/2015 – 12/2015
Department of Surgery, Peking Union Medical College Hospital
- Analysis of Clinical Characteristics and Treatment of Pancreatic Cystic Tumors. (published)

Co-Investigator 09/2014 – 09/2015
Department of Medical Genetics, Peking Union Medical College
- National Student Innovation project
- Tracheal Microenvironment, ANP Metabolism and Airway Tone (published)

Investigator 09/2013 – 08/2014
Department of Physiology, Peking Union Medical College
- National Student Innovation project
- Study of Pathogenic Gene of Primary Hypertrophic Osteoarthropathy (published)

WORK EXPERIENCE

Teaching Assistant 03/2019 – 05/2019
Statistical Methods in Public Health IV
- Mentored students in preparing and performing statistical analysis plan for course projects

Editor-in-Chief 01/2015 – 01/2016
WeChat Easyhin
- Educated public about healthy life style informed by newest research

Founder 05/2015 – 03/2016
AIDS-aid program
- Provided current information to AIDS patients about free medicines and consulted for AIDS patients

Volunteer 01/2012 – 11/2012
Red Cross Society
- Organized blood donation, clothes donation and peer-education on love and relationships

Volunteer/Teacher's Aide 03/2010-08/2010
- Supported education and taught in Taijing primary school in rural Beijing by editing textbooks and giving English lessons twice a month

PUBLICATIONS

1. **Zhang W**, Wang T, Huang S, Zhao X: Identification of a HPGD mutation in three families affected with primary hypertrophic osteoarthropathy. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2018 Apr 10;35(2):156-159

2. Li J*, **Zhang W***, Wang X*, Ma C: Functional magnetic resonance imaging reveals differences in brain activation in response to thermal stimuli in diabetic patients with and without diabetic peripheral neuropathy. PLoS One. 2018 Jan 5;13(1):e0190699. (*co-authors.)

3. Qiu WY*, Yang Q*, **Zhang W***, Wang N, Zhang D, Huang Y, Ma C: The Correlations between Postmortem Brain Pathologies and Cognitive Dysfunction in Aging and Alzheimer's Disease. Curr Alzheimer Res. 2018 Mar 14;15(5):462-473. (*co-authors.)

4. Qiu W, Guo X, Lin X, Yang Q, **Zhang W**, Zhang Y, Zuo L, Zhu Y, Li CR, Ma C, Luo X: Transcriptome-wide piRNA profiling in human brains of Alzheimer's disease. Neurobiol Aging. 2017 Sep;57:170-177.

5. Wang Q*, Jiang K*, **Zhang W***, Qiu W, Li Y, Zheng Y, Wang C, Cao J: Tracheal microenvironment, anp metabolism and airway tone. Chin J Cancer Res. 2016 Oct; 28(5): 519-
527. (*co-authors.)

6. Yang Q, Chen K, Zhang H, **Zhang W**, Gong C, Zhang Q, Liu P, Sun T, Xu Y, Qian X, Qiu W, Ma C. Correlations Between Single Nucleotide Polymorphisms, Cognitive Dysfunction, and Postmortem Brain Pathology in Alzheimer's Disease Among Han Chinese. Neurosci Bull. 2019 Apr;35(2):193-204.

7. You L, Xiao J, Cao Z, **Zhang W**, Liao Q, Dai M, Zhang T, Zhao Y: Analysis of clinical characteristics and treatment of pancreatic cystic tumors. Sci Bull (Beijing). 2016; 61(20): 1551- 1554.

8. Yuan T, Li J, Shen L, **Zhang W**, Wang T, Xu Y, Zhu J, Huang Y, Ma C: Assessment of itch and pain in animal models and human subjects. Adv Exp Med Biol. 2016; 904:1-22.

PROFESSIONAL DEVELOPMENT

Certificate: International First-Aid; Command line tools for genomic data science
Technical Skills: Proficient in STATA programming, familiar with SAS and R programming