Detecting Gene-Gene Interactions for Cleft Lip with/without Cleft Palate in Targeted Sequencing Data

by
Yanzi Xiao

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

Baltimore, Maryland
May 2015

Abstract

Background

Non-syndromic cleft lip with or without cleft palate (NSCL/P) is the most common craniofacial birth defect in humans, affecting 1 in 700 live births. This malformation has a complex etiology where multiple genes and several environmental factors influence risk. At least a dozen different genes have been confirmed to be associated with risk of NSCL/P in previous studies. All the known genetic risk factors cannot fully explain the observed heritability of NSCL/P, and several authors have suggested gene-gene (GxG) interaction may be important in the etiology of this complex and heterogeneous malformation.

Objective

We aimed to detect gene-gene interactions for cleft lip with/without cleft palate in targeted sequencing data.

Methods

We used targeted sequence data on 13 regions identified by previous studies spanning 6.3 MB of the genome in a study of 1,498 case-parent trios. We used R-package Trio to perform a likelihood ratio test (LRT) to test for GxG interaction in both a 1 df test and a 4 df test. To adjust for multiple testing, permutation test was performed to generate empiric p-values.

Results

The most significant 4df LRT was seen with rs6029315 in *MAFB* and rs6681255 in *IRF6* ($p=3.8\times10^{-8}$) in the European group, which remained significant ($p=0.02$) after correcting for multiple comparison via permutation tests. Only 2% of replicates generated under the null hypothesis exceeded this observed test statistic. However, we found no pairwise interaction yielding an empirical $p<0.05$ in the Asian trio group.

Conclusions

Our results suggest that there is statistical GxG interaction between *IRF6* and *MAFB* in the European population. Because *IRF6* is the only gene that has shown consistency across different types of genetic studies, evidence of statistical interaction between markers in/near the genes *IRF6* and *MAFB* is especially interesting.

Acknowledgements

- I would like to express my deep and sincere gratitude to my advisor, Dr. Terri H. Beaty for the continuous support of my study and research, for her patience, motivation, enthusiasm.
- I warmly thank Dr. Ingo Ruczinski for his valuable statistical advice and friendly help.
- My warm thanks are due to Dr. Margaret Taub, Dr. Ferdouse Begum and Ms. Jackie Bidinger. Their kind support and guidance have been of great value in this study.
- I wish to thank Meg Parker, Weiyan Li and Mengdong He for being super awesome!
- Finally, thank you to my parents as well as my dear aunt for all your unconditional love.

Table of Contents

1. BACKGROUND

1.1 Orofacial Clefts

1.1.1 Development Pathogenesis

Orofacial clefts (OFCs) are birth defects in which there are gaps in the normal orofacial structures of the face and mouth caused by abnormal development during the early period of pregnancy. This group of birth defects is the most common craniofacial birth defect among humans affecting 1.7 per 1000 live births [1]. OFCs include three distinct anatomical defects: cleft lip (CL), cleft palate (CP) and cleft lip and palate (CLP). Since CL and CLP share a defect of the primary palate, OFCs can be generally divided into two groups, cleft palate (CP) and cleft lip with or without cleft palate (CL/P).The majority of OFCs cases are considered to be "non-syndromic" which occur as isolated anomaly with no other apparent cognitive or structural abnormality in the child. According to a paper published by Jugessur et al., 70% of all CL/P cases and 50% of all CP cases are considered to be non-syndromic [2].

CL/P and isolated cleft palate (CP) have different developmental pathogenesis. By week 4 of human embryonic development, the frontonasal prominence, paired maxillary processes and paired mandibular processes are formed. During week 5, paired medial and lateral nasal processes come into place. By the end of week 6, the medial nasal processes have merged with maxillary processes to form the upper lip and primary palate. Any disruption of growth during this period could lead to failure of fusion and result in CL/P. On the other hand, CP occurs when there is a disruption in the formation of the secondary palate. The secondary palate constitutes both the floor of the nose and the roof of the mouth. It starts to develop during week 6 of human embryonic development with bilateral outgrowths from the maxillary processes which grow down on either side of the tongue and become the palatal shelves. Later the palate shelves elevate to a horizontal position above the tongue and fuse to form the palate, a process which is completed by week 12 [3].

1.1.2 Descriptive Epidemiology of Orofacial Clefts

According to a report from WHO in 2001, the overall prevalence of orofacial clefts is 1 in 700 live births. The prevalence of CL/P is 3.4-22.9 per 10,000 live births. For CP, the prevalence is 1.3-25.3 per 10,000 live births [4]. There are substantial differences in prevalence of CL/P across racial groups and populations: Asians and Native Americans have the highest rate of 2 per 1,000 live births, Caucasians have a prevalence of 1 per 1,000 and Africa populations have the lowest prevalence rate of 1 per 2,500 live births [5]-[6]. Gender is also shown to be related to orofacial clefts, CL/P is more common in males with a 2:1 ratio of males: females, while CP is twice as frequent in females [6].

1.1.3 Genetic Studies of Orofacial Clefts

1.1.3.1 Family studies and twin studies

Family studies and twin studies have consistently shown that there is a strong genetic component to the etiology of CL/P and CP. The frequency of a positive family history of CL/P (17.3%) was much higher than the prevalence among the relatives of controls (0.5%) in South American populations [7]. A study using data from medical birth registry in Norway showed the recurrence rate of CP among first degree relatives of CP cases was 56 times greater (95% CI =37.2-84.8) than the general population [8]. A twin study from Denmark showed the probandwise concordance rate for CL/P was 50% among monozygotic twins compared with 8% among dizygotic twins. For CP, this probandwise concordance rate was 33% among monozygotic twins compared with 7% among dizygotic twins [9].

1.1.3.2 Linkage studies

Genome-wide linkage studies have suggested several genes are likely to play a causal role in CL/P, but due to modest numbers of multiplex cleft families and their limited size, only a few linkage studies reached genome-wide significance for linkage. In a meta-analysis of 13 genome linkage scans in six populations, the first genome-wide significance results for CL/P were observed in regions 1q32, 2p, 3q27-28, 9q21, 14q21-24 and 16q24 [10]-[11]. This high level of locus heterogeneity, where different families show evidence of linkage to different regions of the genome, argues that multiple genes can lead to orofacial clefts. Subsequent fine-mapping of the 9q21 region identified *FOXE1* as the causative gene in this region [10].

1.1.3.3 Genome-wide association studies

In recent years, genome-wide association studies (GWAS) have been very successful in identifying multiple loci associated with CL/P. To date there have been four GWAS of CL/P [12]-[15] and one for CP [16]. The first successful GWAS, conducted by Birnbaum et al., [12] found extremely strong association between markers in *8q24* and CL/P. The study also confirmed *IRF6* which had prior positive candidate gene and linkage analysis results to be associated with CL/P [12]. The finding in the first GWAS was subsequently replicated in a second GWAS by Grant et al. [13]. In the third GWAS, Mangold et al. identified additional signal near *VAX1* and *NOG* [15]. Unlike other the first three GWAS, the fourth GWAS performed by Beaty et al. utilized a case-parent trio design. The study confirmed previous associations in *8q24* and *IRF6,* and identified novel loci near *MAFB* and *ABCA4* [14]. So far there is only one GWAS of CP which found no genome-wide significant signal, but found some evidence of gene-environment interaction [16].

1.1.3.4 Gene-environment interaction

Marginal gene effects or environmental effects alone may not be apparent when there is interaction between the two. Thus gene-environment interaction studies are important because they aim to describe how genetic and environmental factors could jointly influence the risk of developing disease. In a study conducted by Beaty et al. in 2011, 550 CP case-parent trios were used to test for marginal gene effects, but no SNP achieved genome-wide significance when considered alone. However, there was significant evidence of gene-environment interaction when the model was expanded to consider GxE interaction. There was apparent GxE interaction between *MLLT3* and *SMC2* on chromosome 9 with alcohol consumption, *TBK1* on chromosome 12 and *ZNF236* on chromosome 18 with maternal smoking, and *BAALC* on chromosome 8 with multivitamin supplementation [16].

1.1.3.5 Gene-gene interaction

Despite successfully identifying several different genetic risk factors for CL/P, these polymorphic markers cannot fully explain the observed heritability of CL/P, and several authors have suggested gene-gene (GxG) interaction may be important in the etiology of this complex and heterogeneous birth defect. One study by Li et al. investigated GxG interaction using the same CL/P case-parent trios in the GWAS by Beaty et al. found robust evidence of GxG interaction between markers in *WNT5B* and *MAFB* among Asian and European case-parent trios. Additional evidence of GxG interaction between markers in *WNT5A* and *IRF6* in Asian trios, and markers in the *8q24* region and *WNT5B* in European trios was also found [17].

1.2 Gene-gene interaction

Since the first GWAS was conducted in 2005 [18], a substantial number of genetic risk variants had been discovered. However, most of variants achieving genome-wide significance have a small effect size, and can only explain a small proportion of the overall heritability. There is increasing interest in considering the possibility of GxG interaction, also known as epistasis, which may play an important role in explaining the missing heritability in complex diseases.

1.2.1 Definition of gene-gene interaction/epistasis

The term “epistasis” was first used by William Bateson in 1909 to describe the masking effect whereby a variant/allele at one locus prevents the variant in another locus from manifesting its effect, thereby resulting in deviation from Mendelian inheritance [19].

Coat color variation in Labrador retrievers is a perfect example of epistasis which illustrates the effect of gene-gene interaction on phenotype. The first gene of interest is tyrosinase-related protein (*TYRP1*) gene, which determines the density of the coat's eumelanin pigment granules: dense eumelanin granules result in a black coat, while sparse granules give a chocolate coat color. The second gene of interest is the melanocortin receptor (*MCIR*) gene which determines whether eumelanin is produced at all. The loss of function mutation at this E allele at the *MCIR* cause the coat color to be yellow because pheomelanin is produced rather than eumelanin. In other words, the Labrador is yellow if it is an ‘ee’ homozygote irrespective of the genotype at the B locus at *TYRP1.* However, if the dog carries at least one E allele, only the B locus determines the coat color. In this case, if the dog is a ‘bb’ homozygote it will have a brown coat color; if the dog is heterozygous ’Bb’ the coat color will be black. Thus, the effect B locus on coat color is suppressed if the dog is homozygous ’ee’ at the MCIR locus [20].

Compared to Bateson’s definition of epistasis, a broader definition of epistasis was introduced by Fisher in 1918. Fisher used the term “statistical interaction” to refer to a deviation from additivity in the effect of alleles at two different loci [21]. In other words, whenever the joint effect of two or more genes on a quantitative phenotype cannot be predicted by the sum of their separate effects, then statistical interaction exists. One thing to keep in mind when modelling a statistical interaction is the scale of choice becomes important. Factors that are additive on one scale might show false evidence of interaction under a different scale [22].

1.2.2 Statistical models of epistasis

In 2002, Cordell proposed a generalized linear model to detect epistasis [23]. The saturated or full model is written as

$$logit(p)=\alpha+\beta_1(X_{Aa})+\beta_2(X_{AA})+\gamma_1(X_{Bb})+\gamma_2(X_{BB})+i_{11}(X_{AA}X_{BB})+i_{12}(X_{AA}X_{Bb})+i_{21}(X_{Aa}X_{BB})+i_{22}(X_{Aa}X_{Bb}).$$

This model has a total of nine parameters: α represents the baseline log-odds for an individual who has genotype aabb, where a and b are the respective reference allele at the A and B loci, parameters β_1 and β_2 represent the effect of having one or both A allele at locus A; parameters γ_1 and γ_2 represent the effect of having one or both B allele at locus B, respectively. Effects of GxG interaction are determined by four interaction parameters (i_{11}, $i_{12,}$ i_{21} and i_{22}). If there is no epistasis whatsoever, all of these interaction coefficients are zero ($i_{11}= i_{12}= i_{21}= i_{22}=0$). In this situation, the most appropriate model becomes

$$logit\ (p) = \alpha + \beta_1(X_{Aa}) + \beta_2(X_{AA}) + \gamma_1(X_{Bb}) + \gamma_2(X_{BB}).$$

A 4 degree of freedom (df) likelihood ratio test (LRT) comparing the full model to the model with no interaction parameters can be carried out to test for significant interaction effects.

Since we don't need to assume a model of inheritance (e.g. dominant or recessive inheritance), the full model should give the best fit. However, there are nine parameters in the full model that need to be estimated, which can lead to sparse contingency tables with many empty cells, especially for low frequency variants. Thus, in some situations a model with fewer parameters is preferable. A series of simpler models can be used. For example, we can construct a model that assumes alleles act additively at both A and B loci with only one interaction term:

$$logit\ (p) = \alpha + \beta_1(X_A) + \gamma_1(X_B) + i(X_A\,X_B).$$

Here genotypes are coded as $X_A=$ and $X_B=0$, 1, 2 reflecting the number of risk alleles at the each of the A and B loci, respectively, and i is the single interaction term testing for deviation from complete pairwise additivity. The LRT comparing this model to a reduced model with no interaction (i.e. i=0) may be tested using a LRT with 1 df. Other simplified models include recessive and dominant interaction models. In a recessive interaction model, X_A=0 when the A locus genotype is Aa or aa; $X_A = 1$ when the genotype at locus A is AA, and the same coding schemes is applied for genotypes at the B locus. When considering a dominant interaction model, X_A=0 when genotype at locus A is aa; $X_A = 1$ when the genotype at locus A is AA or Aa, and similarly locus B can use the same coding scheme. Different combinations of specific coding schemes are possible, for example an additive-dominant interaction model. All of these simplified models include a single interaction parameter *i*.

Although population based study designs such as case-control designs are more commonly implemented when testing for GxG interaction, family-based study designs can address these same questions and are more robust against population stratification. The case-parent trio design is the most common family-based study design, and the basic idea behind this study design is to generate "pseudo-controls" using the parent's untransmitted alleles, thus creating a matched case-control design where the observed case is compared to all possible genotypic combinations that could have arisen from the parental mating type. For any single variant, there are three alternative genotypes for pseudo-controls that could have been transmitted to the case, thus the case: pseudo-controls ratio is 1:3 in a conditional logistic regression model. When considering two variants at two independent loci, the case: pseudo-controls ratio in conditional logistic regression becomes 1:15 [24]-[28].

1.2.3 Biological interpretation of epistasis

Whether statistical interaction can suggest biological or functional interaction has been much debated in the field. One problem is that there are different definitions for the commonly used term ‘epistasis’. According to Phillips in a recent review, epistasis can be classified into three types: functional epistasis, compositional epistasis and statistical epistasis [29]. ‘Functional epistasis’ is the interaction of different proteins; ‘compositional epistasis’ harks back to Bateson’s original definition where one allele is blocked by another allele at a different locus, and ‘statistical interaction’ represents the deviation from additivity in the effect of alleles at different loci. It is hard to determine whether statistical evidence of a GxG interaction discovered in conventional statistical models has actual biological meaning, so caution must be used when interpreting statistical evidence for a GxG interaction.

1.3 Research Hypothesis

In this study, we hypothesize there are may be GxG interactions among polymorphic variants identified by targeted sequencing of 13 candidate regions (*8q24, ARHGAP29, BMP4, FGFR2, FOXE1, IRF6, MAFB, MSX1, NOG, NTN1, PAX7, PTCH1, VAX1*) available in the targeted sequencing study described by Leslie et al. [30]. These 13 regions were previously shown to be associated with non-syndromic cleft lip with or without cleft palate (NSCL/P) in either previous GWAS or genome-wide linkage studies.

2. Subjects and Methods

2.1 Study Population

A total of 1,498 cleft case-parent trios were recruited from different sites in China, the Philippines, the United States and Europe and were used for targeted sequencing of 13 genes and regions considered to be prime candidates for containing genes or regulatory elements important in controlling risk to oral clefts (Table 1). After quality control, 1,409 case-parent trios remained available for analyses. Some of these 1,409 case-parent trios were included in a GWAS study [14], but this targeted sequencing study included additional trios. In that previous GWAS, principal components analysis (PCA) was conducted and showed Asian individuals and European and European Americans formed genetically distinct clusters. Therefore, we stratified our data into two groups: an Asian group which contained Filipino and Chinese families (1034 trios), and an European group composed of European and European American families (375 trios).

2.2 Selection of target sequencing regions

We analyzed target sequence data for 13 specific regions (*8q24, ARHGAP29, BMP4, FGFR2, FOXE1, IRF6, MAFB, MSX1, NOG, NTN1, PAX7, PTCH1, VAX1*) spanning 6.3 MB of the genome (Table 2). All 13 regions were identified by previous studies to be associated with oral clefts through GWAS or linkage studies. Nine regions were previously identified by GWAS and/or genome-wide linkage studies and four regions were selected from candidate gene studies.

2.3 Sequencing

According to the manufacturer's protocol (Illumina Inc., San Diego, CA), 1μg of native genomic DNA were used to construct Illumina multiplexed libraries. Reads were mapped to the GRCh37-lite reference sequence using bwa v0.5.9 [31] with the following parameters: -t 4 –q4. Picard (v1.46) was used to merge alignments and mark duplicates. Polymutt (v0.11) was used to perform germline and *de novo* variant calling. Polymutt uses a likelihood-based method considering the parents' genotype information when call *de novo* variants. We used bam-readcount (v.0.4) to identify and flag potential artifact variants if they failed the criteria listed in Table 3. The SNV variant calls were combined into a VCF file. Individual variants with a depth (DP) less than 7 or genotype quality (GQ) less than 20 were removed. Variants located within 75bp of indels or dinucleotide polymorphisms occurring in more than 5% of samples, were included in analyses but were flagged as potential artifacts.

2.4 Family-relationship testing

To evaluate the family relationship between members of these case-parent trios, we used BEAGLE's fast-IBD to calculate identity by descent (IBD) between parents and their offspring. If a parent-child pair shared less than 40% of the targeted region, the trio was dropped from all analysis.

2.5 Selection of common variants and additional quality control

Prior to conducting statistical analyses to detect GxG interaction, we selected common variants and applied additional quality control measures. To increase the power to detect GxG interaction, we only selected SNVs with a minor allele frequency (MAF) larger than 0.2. We also excluded all SNVs with a missing genotype rate larger than 1%. We then tested for Hardy-Weinberg equilibrium in parents within the Asian and European groups separately, and excluded SNVs yielding a HWE $p<1\times10^{-5}$. We used Haploview 4.2 [32] to choose tagging SNVs (defined as $r^2>0.8$) within the Asian and European groups separately.

2.6 Screening step: 1 df Likelihood Ratio Test for GxG interaction

In this study, we implemented an efficient screening strategy to screen all pairwise combinations between common SNVs in these 13 regions using the 1 df likelihood ratio test (LRT) for interaction. All analyses were done using the trio R package [33]. Assuming an additive model for marginal effects of each of two genes, a conditional logistic regression model containing one parameter for each SNP and one parameter for a common interaction term between these two SNPs was fitted and a 1df LRT was performed. The interaction model that incorporated interacting coefficients between two SNVs can be written as

$$logit\ (p) = \alpha + \beta_1\ (X_A) + \gamma_1 (X_B) + i(X_A\ X_B),$$

where X_A and X_B =0,1,2 are coded genotypes at the A and B loci, respectively, reflecting the number of risk allele at that locus. This model has only one interaction parameter and is considerably simpler compared to the full model proposed by Cordell [23] and described above (Section 1.2.2). The simplified model has fewer parameters, making it more efficient to screen for epistasis between all possible SNV pairs. We constructed quantile-quantile (Q-Q) plots by plotting our observed p-values against the expected values under the null distribution. The 95% confidence interval band for this Q-Q plot was obtained under the null hypothesis of no interaction, which should follow a χ^2 distribution with one degree of freedom.

2.7 4 df Likelihood Ratio Test for GxG interaction

For each pairwise combination of the 13 genes/regions, we selected the top 500 most significant pairs of markers from all pairwise combinations of our sub-selected SNPs under this 1 df LRT, and then fit the more general model to create the 4 df interaction model proposed by Cordell [23]. This complete model can be written as

$$logit(p)=\alpha+\beta_1(X_{Aa})+\beta_2(X_{AA})+\gamma_1(X_{Bb})+\gamma_2(X_{BB})+i_{11}(X_{AA}X_{BB})+i_{12}(X_{AA}X_{Bb})+i_{21}(X_{Aa}X_{BB})$$

$+i_{22}\ (X_{Aa}X_{Bb})$ where the coefficients α, β_1, β_2, γ_1 and γ_2 represents the mean effect, additive effect and dominance effect at each the 2 loci A and B. As mentioned above, there are four parameters (i_{11}, i_{12}, i_{21}, i_{22}) representing epistasis effects for all genotypic combinations. We then performed a 4 df likelihood ratio test to comparing the log-likelihood of this full model listed above to that of the null model with no GxG interaction whatsoever, $logit\ (p) = \alpha+ \beta_1(X_{Aa})+ \beta_2(X_{AA})+ \gamma_1(X_{Bb})+ \gamma_2(X_{BB})$.

2.8 Permutation test

The principle behind permutation tests is to use the observed data to simulate the distribution of test statistics under the null hypothesis, and then compare the observed values to this null distribution to obtain an empirical p-value which should be more robust than traditional p-values based on asymptotic assumptions. To perform permutation tests in the context of testing for GxG interaction, we created data sets of pseudo-control children using phased haplotype data for all parents. Haplotype phasing of all parents was done using BEAGLE [34]. In our phased data set, we had one transmitted haplotype and one un-transmitted haplotype for each parent. We created simulated children for each permutation data set by randomly choosing haplotypes from both parents as the transmitted haplotypes, giving a simulated child and three simulated pseudo-controls. We then ran the same analysis on our simulated case: pseudo-controls matched sets. This procedure was repeated 100 times, and we then plotted the maximum test statistic over these 100 replicates to create a distribution of maximum test statistics of the 4 df interaction test expected under the null hypothesis. The empirical p-value can then be calculated by taking the ranking of the observed test statistic among the all test statistics generated over the 100 replicates.

3. Results

After applying quality control filters to our common SNVs from the targeted sequencing data, as described in the Methods section, we were left with 1,075 SNVs and 1,016 SNVs in Europeans and Asians, respectively. We focused on pairwise GxG interactions between different genes/regions in this targeted sequencing data, thus the 13 regions created 78 different gene-gene combinations. To reduce the number of tests in our analysis, we relied on an efficient screening process by performing the 1 df interaction tests for GxG interaction in all pairs of markers between different regions. Figure 1 shows the quantile-quantile (QQ) plots for an exhaustive search of pairwise GxG interaction between markers in *MAFB* & *IRF6* genes using the 1 df interaction test on 375 case-parent trios of European ancestry. The shaded region in these QQ plots corresponds to the 95% concentration band obtained under the null hypothesis of no interaction. We observed an excess of points falling outside the 95% concentration band at the tail of the distribution.

Based on this screening test under the 1 df LRT, we then selected the 500 most significant pairs of markers for each GxG combination, and performed the more general 4 df interaction test for GxG interaction. The most significant SNV pairs are listed in Table 4 and Table 5 for each pair of genes among case-parent trios of European and Asian ancestry, respectively.

The LD structure within genes creates dependency between markers, therefore using a Bonferroni correction for multiple testing over all markers would be too conservative, yielding a much lower probability of rejecting the null hypothesis. An effective way to adjust for multiple testing is to perform permutation tests and generate empirical p-values. Figure 2 shows the most significant pair of SNPs was *rs6681355* in *IRF6* and *rs6029315* in *MAFB* (p=3.8×10^{-08}) in the European group, which remained significant (p=0.02) after correcting for multiple comparison via permutation tests. Only 2% of all 100 replicates generated under the null hypothesis exceeded this observed test statistic. Although there were more case-parent trios in the Asian group, we observed no indication of pairwise interaction in this group (i.e. no pair of SNPs yielded an empirical p-value of less than 0.05 among the larger Asian group).

4. Discussion

Compared to other large scale studies searching for evidence of gene-gene (GxG) interactions, our study implemented an efficient screening strategy to screen all pairwise combinations of highly polymorphic SNVs and focused on the most promising pairs of markers. The 4 df interaction model proposed by Cordell [23] is more generalized and could detect a variety of interactions even if the markers or the genes they tag don't display marginal effects. Moreover, to account for the correlation between markers within a region due to LD between SNPs, we performed permutation testing which can control for multiple comparisons more effectively than a Bonferroni correction when data are correlated.

We detected a GxG interaction between markers that are tagging SNPs in and around *IRF6* and *MAFB* (*rs6681255:rs6029315;* empirical *p* =0.02) in the 375 trios in the European group. Our evidence of statistical interaction between SNPs in *IRF6* and *MAFB* is especially interesting, because *IRF6* is the only gene that has shown consistency across different types of genetic studies, having been identified as the region harboring causal genes for Van der Woude syndrome which is the most common form of syndromic clefting accounting for 2% of all CL/P cases [35], and showing consistent evidence of association with apparently non-syndromic oral clefts [36]. This association finding was subsequently confirmed in a candidate gene study using subjects from several different populations [36]. Genome-wide linkage studies

[10]-[11] and GWAS were also able to replicate evidence for association between polymorphic variants in and near IRF6 and risk of apparently non-syndromic oral clefts [12]-[14]. Animal studies have also shown *IRF6* is expressed in the ectoderm covering the facial processes during their fusion to form the upper lip and primary palate in both mouse and chick [37].

A GWAS study by Beaty et al. (2010) identified several markers near *MAFB* as associated with and linked to an unobserved gene causing CL/P. Expression studies in the mouse also support some role for *MAFB* in palatal development. Sequencing of the *MAFB* exon identified a rare variant (H131Q) which was over represented among Filipino cases [14], although this rare variant seems unlikely to account for the statistical evidence found in the GWAS.

We failed to detect significant GxG interaction in the Asian group of case-parent trios, despite the larger sample size. Many factors could limit our ability to detect GxG interaction between these same SNPs in this larger Asian group. Although we have a large dataset of 1,034 Asian trios, due to different minor allele frequencies between ancestral groups some genotypes might be under represented in Asian populations, making it hard to fit a 4 df interaction model for GxG interaction. In our study, we used tagging SNPs to reduce the number of multiple comparisons and save computer power, however, by relying on highly polymorphic tagging SNPs, we risk pruning out variants critical to identifying GxG interaction.

One of the limitations of our study was its modest sample size and low power to detect GxG interaction. Compared to detecting a marginal effect for any single marker, detecting pairwise or two-way GxG interactions requires a much larger sample size. Even with a very large dataset, some genotypes could still be under represented, making it hard to fit the 4 df GxG interaction model with a total of nine parameters. According to a study published by Mathieu et al. in 2009, it is almost impossible to detect epistasis for markers with allele frequencies below 0.1, even in large datasets with 2000-3000 individuals [38]. Therefore our approach will only be powerful in detecting GxG interaction between highly polymorphic, common SNPs. Another limitation of our study is that we only used parametric logistic regression models to detect GxG interaction. A major challenge of using traditional regression models to detect interaction is specifying the full and reduced models. Additionally, analyzing high-dimensional data which often contains many potential interacting predictor variables could lead to very sparse contingency tables with many empty cells. Machine-learning or data-mining methods represent an alternative approach that do not rely solely a pre-specified model. Limited computer power is another issue we had to consider in our study, although we used tagging SNPs and implemented this efficient two-stage screening strategy, the number of tests is still very large. It took more than 24 hours to run through the analysis plan described here, thus if we were to perform 1000 permutations, even on large CPU clusters it could take 1000 days to complete. Given the limited computer power, we only did 100 permutation tests permutation tests to generate empirical p-values. Finally, the scope of our analysis was limited to targeted sequencing data on 13 regions previously shown by other studies to be associated with CL/P. Variants in regions not showing prior evidence of association, i.e. those without significant marginal effects could also involve significant GxG interaction, but our study was limited to candidate regions that are mostly strongly associated with CL/P, so we might miss some important GxG interactions.

In conclusion, we found some evidence of significant GxG interaction between polymorphic markers in the *IFR6* and *MAFB* genes in a group of case-parent trios of European ancestry. Because IRF6 and MAFB have already shown evidence of being associated with CL/P risk, our evidence of statistical interaction between IRF6 and MAFB is especially intriguing and should be explored more thoroughly.

5. Tables

Table 1. Number of case-parent trios available for analysis (after QC) by population

Population	Country	Total Trios
Asian	China	401
	Philippines	633
Asian TOTAL		**1034**
European	USA	266
	Denmark	9
	Hungary	65
	Spain	26
	Turkey	9
European TOTAL		**375**
TOTAL		**1409**

Table 2 Candidate genes or regions sequenced in this study

GWAS			Candidate Gene		
Gene	**Targeted Region (GRCh37)**	**Total (kbp)**	**Gene**	**Targeted Region (GRCh37)**	**Total (kbp)**
IRF6	chr1:209837199-210468406	631.2	***FOXE1***	chr9:100357692-100876841	519.1
MAFB	chr20:38902646-39614513	711.9	***MSX1***	chr4:4825126-4901385	76.3
ARHGAP29	chr1:94324660-95013109	688.4	***BMP4***	ch14:54382690-54445053	62.4
8q24	chr8:129295896-130354946	1059.1	***FGFR2***	chr10:123096374-123498771	402.4
PAX7	chr1:18772300-19208054	435.8	***PTCH1***	chr9:98133647-98413162	279.5
VAX1	chr10:118421625-119167424	745.8			
NTN1	chr17:8755114-9266060	510.9			
NOG	chr17:54402837-54957390	554.6			

Table 3 False positive filters for single nucleotide variants

Filter	**Value**
Maximum difference of mapping quality between variant and reference reads	30
Maximum difference of average supporting read length between variant and reference reads	25
Minimum length of a flanking homopolymer of same base to remove a variant	5
Minimum average relative distance from start/end of read, given as fraction	0.10
Minimum representation of variant allele on each strand	0.01
Minimum number of variant-supporting reads	4
Minimum average relative distance to effective 3prime end of read (real end or Q2) for variant-supporting reads	0.20
Minimum variant allele frequency	0.05

Table 4 Most significant result the 4 df Likelihood Ratio Test for GxG interaction in 375 European case-parent trios

First Gene	Second Gene	Marker 1	Marker 2	Test Statistic	p-value
8q24	ARHGAP29	rs1356762	rs61782236	20.03711	0.000491
8q24	BMP4	rs4236742	rs2224835	13.67618	0.008404
8q24	FGFR2	rs1464154	rs10886946	19.8853	0.000526
8q24	FOXE1	rs72730212	rs16923269	25.3543	4.27E-05
8q24	IRF6	rs4602853	rs28630860	17.69383	0.001416
8q24	MAFB	rs6470670	rs3092775	28.04986	1.22E-05
8q24	MSX1	rs12676542	rs2220746	20.16729	0.000463
8q24	NOG	rs1372992	rs12450049	19.8325	0.000539
8q24	NTN1	rs13265167	rs7207143	21.15418	0.000295
8q24	PAX7	rs13251901	rs4075768	20.0665	0.000485
8q24	PTCH1	rs13249571	rs62558314	24.08714	7.67E-05
8q24	VAX1	rs10090304	rs1681736	22.29485	0.000175
ARHGAP29	BMP4	rs12121974	rs12883570	16.54116	0.002373
ARHGAP29	FGFR2	rs17394161	rs10466213	21.45103	0.000258
ARHGAP29	FOXE1	rs472908	rs2120263	21.10151	0.000302
ARHGAP29	IRF6	rs11165073	rs74487756	20.67322	0.000368
ARHGAP29	MAFB	rs2022395	rs6065286	22.12531	0.000189
ARHGAP29	MSX1	rs4147848	rs730575	29.00544	7.80E-06
ARHGAP29	NOG	rs1761375	rs227688	21.81761	0.000218
ARHGAP29	NTN1	rs1765622	rs7222455	20.51077	0.000396
ARHGAP29	PAX7	rs1320502	rs4920501	22.64593	0.000149
ARHGAP29	PTCH1	rs12088309	rs357542	19.42877	0.000647
ARHGAP29	VAX1	rs762485	rs2921962	23.74542	8.98E-05
BMP4	FGFR2	rs4243595	rs12256320	14.27175	0.006476
BMP4	FOXE1	rs8014363	rs7033765	18.61077	0.000937
BMP4	IRF6	rs2761884	rs1983614	20.17394	0.000461
BMP4	MAFB	rs11157993	rs3092011	19.55044	0.000612
BMP4	MSX1	rs72680512	rs4689186	18.76032	0.000876
BMP4	NOG	rs12587398	rs12951993	17.36542	0.001641
BMP4	NTN1	rs8014363	rs7208881	17.70963	0.001406
BMP4	PAX7	rs8014363	rs1537843	15.16696	0.004367
BMP4	PTCH1	rs11157993	rs11793640	19.27656	0.000693
BMP4	VAX1	rs8014071	rs1638673	18.61467	0.000935
FGFR2	FOXE1	rs1696835	rs10739476	19.90517	0.000521
FGFR2	IRF6	rs2420941	rs845451	20.59975	0.00038
FGFR2	MAFB	rs10886928	rs6016377	20.18951	0.000458
FGFR2	MSX1	rs4752571	rs4435686	19.46764	0.000636

FGFR2	NOG	rs4752571	rs11654202	20.94156	0.000325
FGFR2	NTN1	rs10466213	rs61409745	20.0787	0.000482
FGFR2	PAX7	rs10510099	rs626600	21.58006	0.000243
FGFR2	PTCH1	rs35462105	rs357521	15.74933	0.003375
FGFR2	VAX1	rs34143724	rs11593912	20.6161	0.000377
FOXE1	IRF6	rs12001675	rs633352	17.4501	0.00158
FOXE1	MAFB	rs3780419	rs3092011	20.62923	0.000375
FOXE1	MSX1	rs1475695	rs13117093	20.50011	0.000398
FOXE1	NOG	rs13049	rs8074637	16.67813	0.002232
FOXE1	NTN1	rs10984601	rs7215971	20.93009	0.000327
FOXE1	PAX7	rs12349452	rs61761365	21.46416	0.000256
FOXE1	PTCH1	rs6478391	rs1889617	14.93449	0.004839
FOXE1	VAX1	rs10984977	rs181512	16.48943	0.002428
IRF6	**MAFB**	**rs6681355**	**rs6029315**	**40.25455**	**3.83E-08**
IRF6	MSX1	rs1983614	rs6851263	17.96455	0.001254
IRF6	NOG	rs590152	rs2159226	25.01446	5.00E-05
IRF6	NTN1	rs599021	rs181533	22.89332	0.000133
IRF6	PAX7	rs2484030	rs10907314	25.72278	3.60E-05
IRF6	PTCH1	rs590152	rs357565	18.19713	0.001129
IRF6	VAX1	rs4421592	rs3010467	19.96762	0.000507
MAFB	MSX1	rs6029145	rs6851263	26.7483	2.23E-05
MAFB	NOG	rs4812455	rs10852990	19.65669	0.000584
MAFB	NTN1	rs6029421	rs8081873	32.17376	1.76E-06
MAFB	PAX7	rs6029182	rs11584404	31.25273	2.72E-06
MAFB	PTCH1	rs6102167	rs10990303	21.28513	0.000278
MAFB	VAX1	rs6072087	rs11197835	17.74937	0.001381
MSX1	NOG	rs2933586	rs4605230	18.37558	0.001042
MSX1	NTN1	rs2968669	rs9892906	27.96647	1.27E-05
MSX1	PAX7	rs3815544	rs61760688	18.53108	0.000971
MSX1	PTCH1	rs60726571	rs1932075	17.95906	0.001257
MSX1	VAX1	rs2968702	rs2420309	17.20441	0.001764
NOG	NTN1	rs8074637	rs2315286	29.62187	5.84E-06
NOG	PAX7	rs8073455	rs9439729	25.30977	4.36E-05
NOG	PTCH1	rs3867600	rs357551	15.64858	0.003529
NOG	VAX1	rs8069500	rs10886011	17.49259	0.00155
NTN1	PAX7	rs7219272	rs4075768	20.59273	0.000381
NTN1	PTCH1	rs62069969	rs574688	18.48369	0.000992
NTN1	VAX1	rs7214739	rs2619106	21.65081	0.000235
PAX7	PTCH1	rs28441017	rs28716262	18.7936	0.000863
PAX7	VAX1	rs2236799	rs11197835	25.224	4.54E-05
PTCH1	VAX1	rs4742697	rs77204400	23.84669	8.57E-05

Table 5 Most significant result the 4df Likelihood Ratio Test for GxG interaction in Asians

First Gene	Second Gene	Marker 1	Marker 2	Test Statistic	p-value
8q24	ARHGAP29	rs873232	rs3789398	25.81935	3.44E-05
8q24	BMP4	rs7845615	rs3742556	17.76259	0.001373
8q24	FGFR2	rs1464154	rs11200102	21.98863	0.000201
8q24	FOXE1	rs72609875	rs12352658	20.45956	0.000405
8q24	IRF6	rs10111530	rs6540559	24.94469	5.16E-05
8q24	MAFB	rs1516960	rs7509091	18.57721	0.000951
8q24	MSX1	rs1835851	rs56398386	23.6565	9.36E-05
8q24	NOG	rs9643244	rs4794668	17.26239	0.001719
8q24	NTN1	rs10956419	rs2429370	24.3927	6.66E-05
8q24	PAX7	rs55830016	rs2841087	23.50632	0.0001
8q24	PTCH1	rs9643244	rs28716262	19.81956	0.000542
8q24	VAX1	rs6984251	rs181505	23.87344	8.47E-05
ARHGAP29	BMP4	rs581244	rs67475977	21.20612	0.000288
ARHGAP29	FGFR2	rs3789692	rs2981451	21.68104	0.000232
ARHGAP29	FOXE1	rs582798	rs12347079	20.07933	0.000482
ARHGAP29	IRF6	rs1324214	rs650854	21.46012	0.000257
ARHGAP29	MAFB	rs950283	rs6072160	23.22121	0.000114
ARHGAP29	MSX1	rs4147830	rs3821949	19.44953	0.000641
ARHGAP29	NOG	rs6674226	rs8069500	19.992	0.000501
ARHGAP29	NTN1	rs2022378	rs3785995	20.21006	0.000454
ARHGAP29	PAX7	rs10874810	rs9439697	19.68882	0.000575
ARHGAP29	PTCH1	rs3761910	rs2149722	29.04059	7.67E-06
ARHGAP29	VAX1	rs6698203	rs1630816	20.96604	0.000322
BMP4	FGFR2	rs2738265	rs2936861	27.40754	1.64E-05
BMP4	FOXE1	rs56312905	rs1886002	19.53078	0.000618
BMP4	IRF6	rs4898820	rs968033	17.50576	0.001541
BMP4	MAFB	rs12895262	rs6102096	19.62203	0.000593
BMP4	MSX1	rs12895262	rs6823800	18.17776	0.001139
BMP4	NOG	rs2147105	rs8073799	18.53637	0.000969
BMP4	NTN1	rs4243595	rs12602314	20.83272	0.000342
BMP4	PAX7	rs6572930	rs515739	22.26337	0.000178
BMP4	PTCH1	rs12895971	rs10985356	19.52068	0.000621
BMP4	VAX1	rs1951866	rs877396	18.51047	0.000981
FGFR2	FOXE1	rs9420327	rs58100391	19.47045	0.000635
FGFR2	IRF6	rs11200101	rs12025057	22.97392	0.000128
FGFR2	MAFB	rs2936861	rs6102078	20.26641	0.000442
FGFR2	MSX1	rs1896422	rs2131453	19.60986	0.000596

FGFR2	NOG	rs2936874	rs227725	19.41057	0.000653
FGFR2	NTN1	rs2935693	rs4791823	20.70899	0.000362
FGFR2	PAX7	rs12763463	rs11488726	26.23335	2.84E-05
FGFR2	PTCH1	rs2936864	rs55952687	19.49601	0.000628
FGFR2	VAX1	rs752736	rs1665668	24.30332	6.94E-05
FOXE1	IRF6	rs2417730	rs12083466	18.15918	0.001149
FOXE1	MAFB	rs4743128	rs2024574	22.52515	0.000158
FOXE1	MSX1	rs2417729	rs80227476	19.55435	0.000611
FOXE1	NOG	rs1886002	rs28664662	22.74566	0.000142
FOXE1	NTN1	rs77159549	rs2551799	22.9835	0.000128
FOXE1	PAX7	rs2808685	rs34988159	19.99556	0.0005
FOXE1	PTCH1	rs958346	rs1335048	19.81282	0.000544
FOXE1	VAX1	rs3994138	rs363312	22.10043	0.000191
IRF6	MAFB	rs72649973	rs2866114	17.1373	0.001818
IRF6	MSX1	rs1473683	rs12639983	16.78104	0.002132
IRF6	NOG	rs10863785	rs12450244	18.56595	0.000956
IRF6	NTN1	rs7511737	rs117996464	20.48185	0.000401
IRF6	PAX7	rs12029138	rs2883890	19.86867	0.00053
IRF6	PTCH1	rs1040426	rs16909974	20.90271	0.000331
IRF6	VAX1	rs1883308	rs17095763	22.18989	0.000184
MAFB	MSX1	rs11907397	rs1907980	21.5971	0.000241
MAFB	NOG	rs2425406	rs8069500	23.78411	8.82E-05
MAFB	NTN1	rs13041631	rs72809908	28.16087	1.16E-05
MAFB	PAX7	rs35929622	rs4075768	21.0293	0.000312
MAFB	PTCH1	rs6016400	rs117758836	19.45235	0.00064
MAFB	VAX1	rs10485671	rs10736259	19.38869	0.000659
MSX1	NOG	rs9291153	rs7222986	24.3445	6.81E-05
MSX1	NTN1	rs12532	rs2429370	21.05962	0.000308
MSX1	PAX7	rs4395446	rs2236806	21.56339	0.000245
MSX1	PTCH1	rs74485582	rs34556283	17.7815	0.001362
MSX1	VAX1	rs1907980	rs758367	16.33392	0.002602
NOG	NTN1	rs17821518	rs12452003	28.78746	8.63E-06
NOG	PAX7	rs227723	rs2236832	24.7415	5.67E-05
NOG	PTCH1	rs887088	rs10990355	20.54162	0.00039
NOG	VAX1	rs1816806	rs181505	18.96081	0.0008
NTN1	PAX7	rs12452951	rs6672970	21.36184	0.000268
NTN1	PTCH1	rs57675223	rs10990447	24.18372	7.34E-05
NTN1	VAX1	rs9901367	rs1468539	21.46883	0.000256
PAX7	PTCH1	rs851123	rs28563972	23.42947	0.000104
PAX7	VAX1	rs2223585	rs1638667	22.26777	0.000177
PTCH1	VAX1	rs16909974	rs3125617	21.70604	0.000229

6. Figures

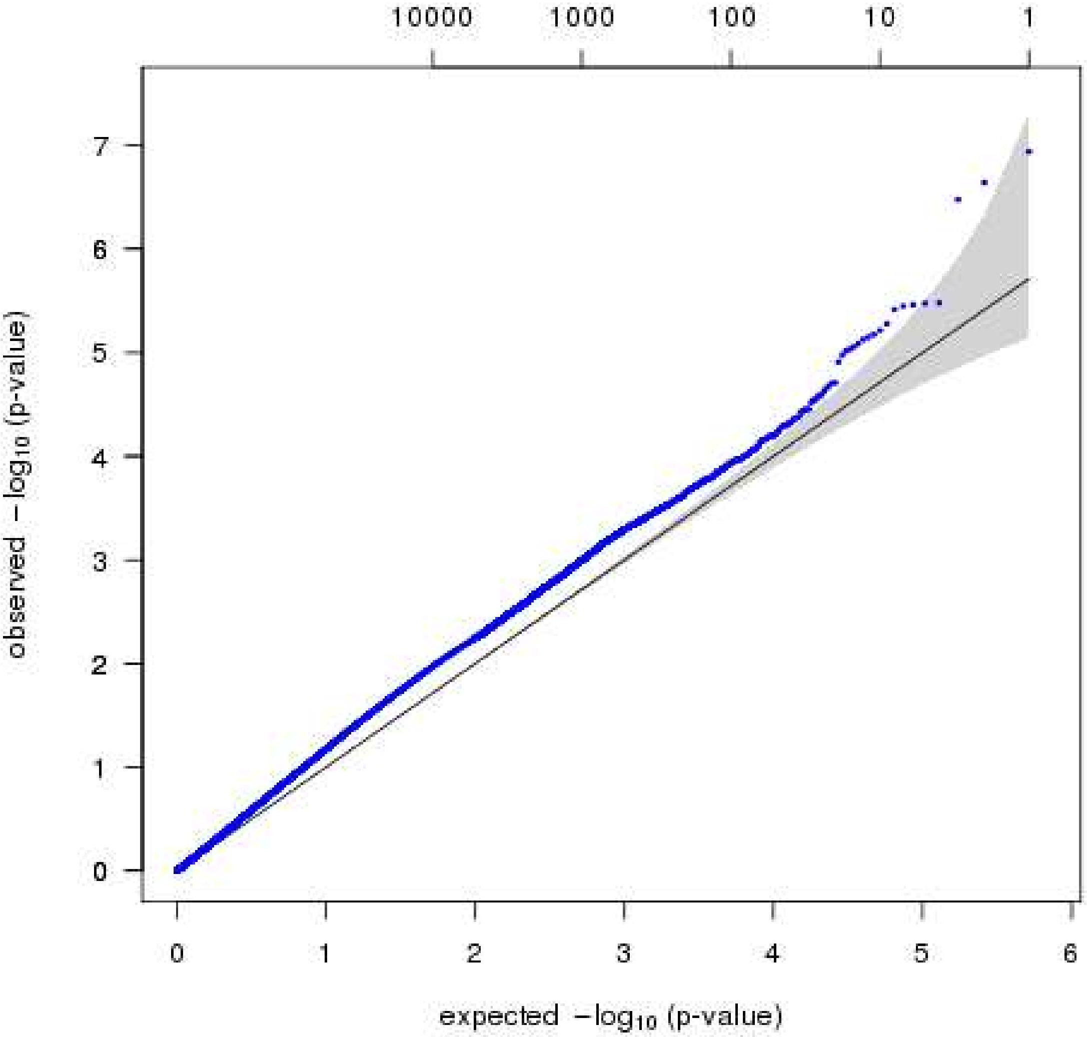

FIGURE 1: QQ plot of exhaustive search of pairwise GxG interaction in *IRF6* & *MAFB* using the 1 df interaction test on 375 case-parent trios of European ancestry

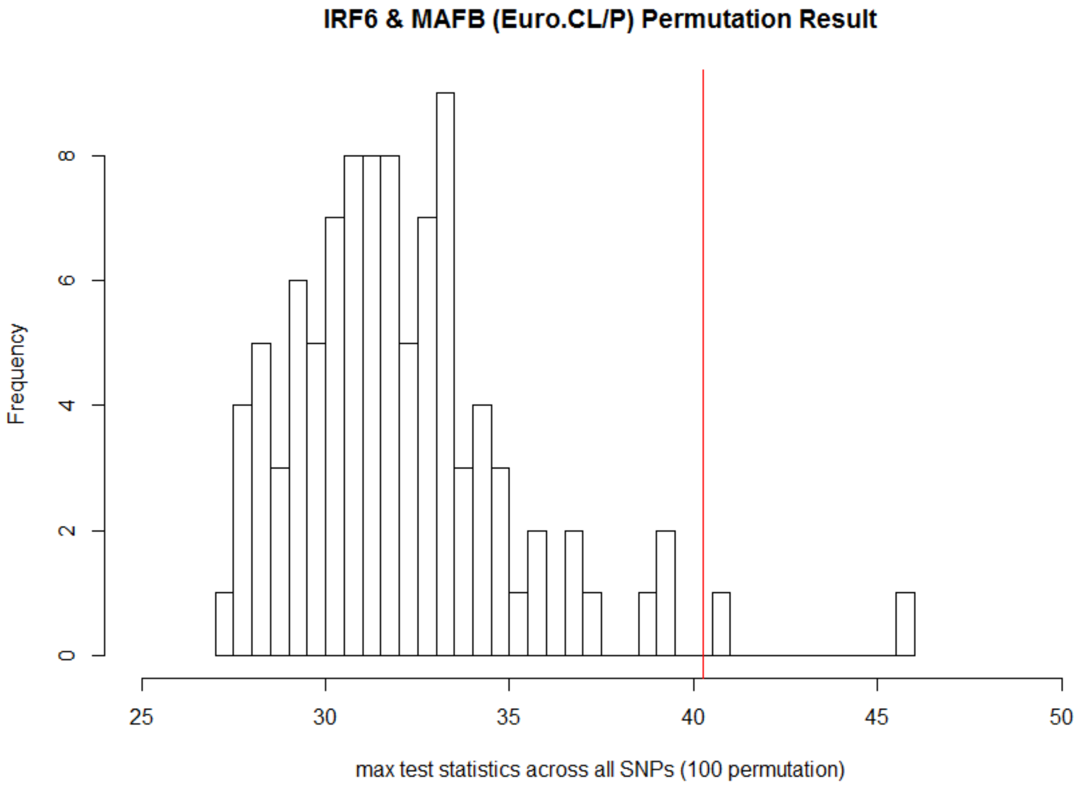

FIGURE 2. Distribution of maximum LRT values over 100 replicates. Histograms represents the frequency of the maximum LRT statistic generated under the null hypothesis of complete independence between markers (i.e. no GxG interaction) for

7. References

1. Rahimov, F., Jugessur, A., & Murray, J. C. (2012). Genetics of nonsyndromic orofacial clefts. *The Cleft Palate-Craniofacial Journal*, *49*(1), 73-91.
2. Jugessur, A., Farlie, P. G., & Kilpatrick, N. (2009). The genetics of isolated orofacial clefts: from genotypes to subphenotypes. *Oral diseases*, *15*(7), 437-453.
3. Marazita, M. L. (2012). The evolution of human genetic studies of cleft lip and cleft palate. *Annual review of genomics and human genetics*, *13*, 263.
4. Mossey, P. A., & Catilla, E. E. (2003). Global registry and database on craniofacial anomalies: Report of a WHO Registry Meeting on Craniofacial Anomalies.
5. Dixon, M. J., Marazita, M. L., Beaty, T. H., & Murray, J. C. (2011). Cleft lip and palate: understanding genetic and environmental influences. *Nature Reviews Genetics*, *12*(3), 167-178.
6. Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., & Shaw, W. C. (2009). Cleft lip and palate. *The Lancet*, *374*(9703), 1773-1785.
7. Menegotto, B. G., & Salzano, F. M. (1991). Clustering of malformations in the families of South American oral cleft neonates. *Journal of medical genetics*,*28*(2), 110-113.
8. Sivertsen, Å., Wilcox, A. J., Skjærven, R., Vindenes, H. A., Åbyholm, F., Harville, E., & Lie, R. T. (2008). Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *Bmj*,*336*(7641), 432-434.
9. Grosen, D., Bille, C., Petersen, I., Skytthe, A., von Bornemann Hjelmborg, J., Pedersen, J. K., ... & Christensen, K. (2011). Risk of oral clefts in twins.*Epidemiology (Cambridge, Mass.)*, *22*(3), 313.
10. Marazita, M. L., Lidral, A. C., Murray, J. C., Field, L. L., Maher, B. S., McHenry, T. G., ... & Arcos-Burgos, M. (2009). Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results.*Human heredity*, *68*(3), 151.
11. Marazita, M. L., Murray, J. C., Lidral, A. C., Arcos-Burgos, M., Cooper, M. E., Goldstein, T., ... & Roddick, L. G. (2004). Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *The American Journal of Human Genetics*, *75*(2), 161-173.
12. Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., ... & Mangold, E. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature genetics*, *41*(4), 473-477.

13. Grant, S. F., Wang, K., Zhang, H., Glaberson, W., Annaiah, K., Kim, C. E., ... & Hakonarson, H. (2009). A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *The Journal of pediatrics*, *155*(6), 909-913.

14. Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., ... & Kirke, P. N. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature genetics*, *42*(6), 525-529.

15. Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., ... & Nöthen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature genetics*, *42*(1), 24-26.

16. Beaty, T. H., Ruczinski, I., Murray, J. C., Marazita, M. L., Munger, R. G., Hetmanski, J. B., ... & Scott, A. F. (2011). Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genetic epidemiology*, *35*(6), 469-478.

17. Li, Q., Kim, Y., Suktitipat, B., Hetmanski, J. B., Marazita, M. L., Duggal, P., ... & Bailey-Wilson, J. E. (2015). Gene-Gene Interaction Among WNT Genes for Oral Cleft in Trios. *Genetic epidemiology*.

18. Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385-389.

19. Bateson, W. (1902). *Mendel's principles of heredity*. University Press.

20. Schmutz, S. M., Berryere, T. G., & Goldfinch, A. D. (2002). TYRP1 and MC1R genotypes and their effects on coat color in dogs. *Mammalian Genome*, *13*(7), 380-387.

21. Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the royal society of Edinburgh*, *52*(02), 399-433.

22. Frankel, W. N., & Schork, N. J. (1996). Who's afraid of epistasis?. *Nature genetics*, *14*(4), 371-373

23. Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, *11*(20), 2463-2468.

24. Cordell, H. J., & Clayton, D. G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *The American Journal of Human Genetics*, *70*(1), 124-141.

25. Cordell, H. J., Barratt, B. J., & Clayton, D. G. (2004). 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. *Genetic epidemiology*, *26*(3), 167-185.

26. Martin, E. R., Ritchie, M. D., Hahn, L., Kang, S., & Moore, J. H. (2006). A novel method to identify gene–gene effects in nuclear families: the MDR-PDT.*Genetic epidemiology*, *30*(2), 111-123.
27. Kotti, S., Bickeböller, H., & Clerget-Darpoux, F. (2007). Strategy for detecting susceptibility genes with weak or no marginal effect. *Human heredity*, *63*(2), 85-92.
28. Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Mangold, J. E., Zhu, J., ... & Li, M. D. (2008). A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *The American Journal of Human Genetics*, *83*(4), 457-467.
29. Phillips, P. C. (2008). Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*(11), 855-867
30. Leslie, E. J., Taub, M. A., Liu, H., Steinberg, K. M., Koboldt, D. C., Zhang, Q., ... & Murray, J. C. (2015). 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. *The American Journal of Human Genetics*, *96*(3), 397-411.
31. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589-595.
32. Barrett, J. C., Fry, B., Maller, J. D. M. J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263-265.
33. Schwender, H., Li, Q., Neumann, C., Taub, M. A., Younkin, S. G., Berger, P., ... & Ruczinski, I. (2014). Detecting Disease Variants in Case-Parent Trio Studies Using the Bioconductor Software Package trio. *Genetic epidemiology*, *38*(6), 516-522.
34. Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, *85*(6), 847-861.
35. Kondo, S., Schutte, B. C., Richardson, R. J., Bjork, B. C., Knight, A. S., Watanabe, Y., ... & Murray, J. C. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature genetics*, *32*(2), 285-289.
36. Zucchero, T. M., Cooper, M. E., Maher, B. S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., ... & Murray, J. C. (2004). Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *New England Journal of Medicine*, *351*(8), 769-780.
37. Knight, A. S., Schutte, B. C., Jiang, R., & Dixon, M. J. (2006). Developmental expression analysis of the mouse and chick orthologues of IRF6: the gene

mutated in Van der Woude syndrome. *Developmental dynamics*, *235*(5), 1441-1447

38. Emily, M., Mailund, T., Hein, J., Schauser, L., & Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, *17*(10), 1231-1240.

8. Curriculum Vitae

Yanzi Xiao

615 N. Wolfe St, Room W6517 | Baltimore, MD
(617) 893-0017 | xiaoyanzijane@gmail.com

EDUCATION

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Master of Health Science (MHS), expected 2015
Concentration: Human Genetics/Genetic Epidemiology
Related Coursework: Methods in Biostatistics I,II,III; Epidemiologic Methods I,II,III IV; Principles of Genetic Epidemiology I,II,III IV; Statistics for Genomics; PERL for Bioinformatics; Analysis of Biological Sequences

Harbin Medical University, Harbin, China
Bachelor of Medicine (MB), 2013

PROFESSIONAL EXPERIENCE

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, May 2014 – Present
Student Investigator

- Master's Thesis: *Detecting Gene-Gene (GXG) interactions for cleft lip with or without cleft palate in targeted sequencing data.*
- Independently conducted replication of Genome-Wide Association Studies (GWAS) in 1,500 case-parent trios of European and Asian ancestry.
- Coordinated a multi-center collaboration assessing gene-environment (GXE) interactions.
- Won 2rd Place at "Hopkins Genetics Research Day Poster Competition".

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, January 2014 – December 2014
Teaching Fellow

- Provided teaching support for graduate level course, "Principles of Genetic Epidemiology II".
- Gave a talk on "How to utilize PLINK: A whole genome data analysis toolset".
- Responsible for advising students, curriculum development, and grading papers.

BESURE Study, Baltimore, MD, September 2014 – Present
Maryland State Certified HIV tester & counsellor

- *BESURE Study is a CDC funded National HIV Behavioral Surveillance Study, in collaboration with the Maryland Department of Health and Mental Hygiene.*
- Perform rapid HIV test and health behavior counselling in LGBT communities.

Center for Disease Control and Prevention, Guangxi, China, June 2012 – August

2012

Intern, AIDS Prevention Center

- Utilized statistical skills to analyze three-year's worth of data collected from two hundred sentinels in 14 cities/counties in Guangxi Province.

Harbin Medical University, Harbin, China, June 2012 – May 2013

Research Assistant, Department of Human Genetics

- Undergraduate thesis *"Family Based Association Study of DOCK4 gene polymorphisms linked to autism in Chinese Han population"* was awarded "Best Thesis of the Year".

Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, June 2011 – May 2012

Medical Intern

- Completed one year of medical training, including all clinical courses and diagnostic training.

SKILLS

Language: Fluent in Mandarin and Cantonese

Computer: R, PERL, SAS, STATA, PLINK, LaTex, UNIX/Linux operating system, MS office