METHODS FOR FAMILY-BASED DESIGNS IN GENETIC EPIDEMIOLOGY STUDIES

by

Jenna Colavincenzo Carlson

BS, California Polytechnic State University, 2012

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Jenna Colavincenzo Carlson

It was defended on

March 20, 2017

and approved by

Dissertation Advisor:
Eleanor Feingold, PhD
Professor
Departments of Biostatistics and Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Members:
Steward J. Anderson, PhD
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Vincent C. Arena, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Wei Chen, PhD
Associate Professor
Departments of Pediatrics, Human Genetics, and Biostatistics
School of Medicine
University of Pittsburgh

Elizabeth J. Leslie, PhD
Assistant Professor
Department of Oral Biology
School of Dental Medicine
University of Pittsburgh

METHODS FOR FAMILY-BASED DESIGNS IN GENETIC EPIDEMIOLOGY STUDIES

Jenna Colavincenzo Carlson, PhD

University of Pittsburgh, 2017

ABSTRACT

In genetic epidemiology studies of complex traits, there are two main design types through which we can study complex traits. The first is population-based, in which independent cases and controls are collected to assess the difference in the underlying genetic makeup between affected and unaffected individuals. The other is family-based, in which data from families with at least one affected individual are collected. This allows for the study of the transmission of genetic variants between parent and offspring and how genetic variants differ between the affected individual(s) and the unaffected individuals within a family.

We examine two hallmarks of complex traits in this dissertation. The first is the combination of mixed data types into a single likelihood, leveraging assumptions about the genotype frequencies to the extent that the data support them. To do this we will employ an empirical Bayes-type shrinkage estimation approach. Combining multiple data structures into a robust joint analysis may provide additional information about the disease loci driving complex traits. Secondly, we will examine heterogeneous presentation of traits associated with complex disorders. This phenotypic heterogeneity may arise due to genetic underpinnings, different environmental exposures, or perhaps by unknown factors. Specifically, we will address the following questions: (1) How can family data be combined with case-control data from the same study to improve estimates of disease association in a way that is robust to model misspecification? (2) How can genetic sources of phenotypic heterogeneity be identified in case-control and family-based studies?

The public health significance of this research is that these methods will further understanding of the genetic architecture and will provide framework for studying other complex traits. Knowing the underlying genetic structure of a complex disease like orofacial clefting will aid in identifying any possible modifiable environmental factors that may also be contributing to the etiology of the disease. In order to identify those factors, we must have foundational knowledge of the biologic mechanism through which OFCs arise.

Keywords: genome-wide association study, empirical Bayes-type estimation, shrinkage estimation, phenotypic heterogeneity, complex traits.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The goal of genetic studies is to further understand the mechanisms contributing to a phenotype by measuring association between genetic variants and the phenotype in some population. Genome-wide association studies (GWAS) aim to do so by assessing the effect of single-nucleotide polymorphisms (SNPs) on a trait statistically. Traditionally, an association test is performed at every genetic marker genome-wide and the markers demonstrating the most statistical significance are considered for further interrogation. These associations identify candidate loci for genetic association with the trait. Importantly, these associations are not necessarily causal, as statistical power is influenced by the allele frequency. Rather, markers implicated in association studies are thought to be in linkage disequilibrium (i.e. correlated) with true causal genetic marker(s).

The power of GWAS to identify a true association between a SNP and trait is dependent on the variability present in the phenotype and how much of that variability can be explained by the SNP [35]. The variability in the phenotype is determined by the effect size of the variant and the allele frequencies in the sample. Because of this, analyzing both rare variants and variants with small effect size can pose problems in GWAS. Additionally, statistical power to detect association between a genetic marker and a phenotype is decreased as phenotypic variation which is not directly attributable to the genetic variant increases. This is common in complex diseases which typically have heterogeneous presentations.

Furthermore, some traits are driven by a few loci with large effect sizes, whereas others are controlled by more genetic loci and numerous factors including admixture, epistasis, and environmental exposures. Investigating complex architectures requires examining population structure and potential allele-frequency differences across populations. Spurious associations can occur for SNPs with varying allele frequencies and trait distributions by population by

population [35]. Additionally, investigation of potential gene-by-gene interactions (i.e. epistasis) and gene-by-environment interactions is warranted. These interactions occur when the effect a genetic locus has on a trait is modified by either another genetic locus (gene-by-gene interaction) or an external environmental factor (gene-by-environment interaction). These complexities in genetic architecture present challenges in GWAS. There are virtually endless possibilities for the underlying genetic model of complex traits. These models can include any of the considerations mentioned above, including but not limited to rare variant contributions, differing effect sizes, population differences, gene-by-gene interactions, and gene-by-environment interactions. Assessing a genetic variant's association with a trait having complex genetic architecture presents an interesting challenge.

There are two primary sampling schemes for GWAS. The first is population-based, comprised of unrelated individuals; the second is family-based, consisting of related individuals. Although these structures are traditionally viewed as separate analyses, they may be combined as mixed data structure and analyzed together. Family-based studies are unique in that they are robust against population stratification; spurious statistical associations due to differences in allele frequencies across populations are generally not discovered in family-based samples [27]. These data structures also allow for the study of transmission of alleles from parent to offspring. In contrast, the traditional epidemiological population-based study design is easy to implement as it does not require recruitment of every individual of interest from within a family. Moreover, population-based designs are more powerful to detect common, weak genetic associations [66].

We will investigate two of the considerations discussed above – examining heterogeneous phenotypes and mixing data structures from population-based and family-based collection methods. Mixed data structures are being used increasingly in the study of complex traits because they offer the advantages of both population-based and family-based designs without limiting the study with the disadvantages that come with selecting only one method [29]. As mentioned previously, phenotypic heterogeneity is a hallmark of complex traits and can reduce the ability to detect true genetic associations. However, there may be genetic differences responsible for the variability in phenotype, the identification of which would further elucidate the genetic underpinnings of complex traits. This dissertation addresses

the philosophical and statistical considerations for mixing data structures and addressing phenotypic heterogeneity.

1.1 MIXED DATA STRUCTURES

Genetic association studies can generally be divided into two main design types – population-based studies and family-based studies. In the study of dichotomous traits, population-based case-control designs, which directly compare the frequency of genetic variants between (usually independent) cases and controls, are widely used for association studies. The goal of these studies is to identify potential genetic loci with differential frequency between cases and controls which may correspond to conferring disease risk. Case-control designs are increasingly being used for GWAS due to the ease in recruitment and the decreasing cost of genotyping large numbers of individuals [11] [36] [65].

Alternatives to case-control designs include various family-based designs, including the case-parent trio design. The most common analysis technique with case-parent trios is arguably the transmission/disequilibrium test (TDT). The TDT examines case-parent trios in which the proband is an incident case [71]. In this situation, the allele at each locus of interest (or genome-wide) is tested for whether the transmission of that allele from parent to offspring is different from what is expected under Mendelian inheritance (i.e. each allele has a 50% chance of being transmitted). This would provide evidence that cases are under/over enriched for an allele due to the increased/decreased risk harbored by that variant.

The case-parent trio design is robust against population stratification as the methods for analyzing such data include some form of conditioning on parental genotypes, which eliminates potential bias from differing genetic background. In this setting, studying parents provides perfectly matched controls for each incident case, and thus is robust to any existing population substructure. However, case-parent trios are often difficult to collect as they require the ascertainment of both DNA specimens and phenotyping for each member of the trio. Moreover, the cost of genotyping trios is three times that for each case or control without a corresponding linear increase in statistical power.

While population-based studies have increased statistical power over the family-based designs of the same number of individuals, association signals detected from this method may be due to uncontrolled confounding factors. In particular, case-control designs are susceptible to confounding population stratification in which the genetic ancestry is associated with both allele frequency and disease incidence.

Population-based data collection is frequently combined with family-based data collection for many reasons. First, in an effort to gather as much information about a trait as possible, genotypic and phenotypic information is often collected on every available person. Secondly, and arguably most importantly, combining family-based data with population-based data protects against false positive association due to population substructure. Thus, combining these approaches offers increased statistical power and protection against false positives.

An overview of the basic analyses for case-control and case-parent trio designs and a brief review of the available methods for combining these data together is given in chapter 2.

1.2 PHENOTYPIC HETEROGENEITY IN COMPLEX TRAITS

The model of Mendelian inheritance offers a simple explanation of the genetic architecture of a trait. It prescribes that a single gene locus produces the trait in either recessive or dominant pattern in families. However, many traits do not follow such a straightforward model of genetic architecture.

Complex traits are those that do not exhibit classic Mendelian recessive or dominant inheritance attributable to a single gene locus [36]. Any break in a direct genotype-phenotype association (i.e. the same genotype resulting in different phenotypes, or different genotypes resulting in the same phenotype) increases the genetic complexity of the trait. This can be caused by numerous factors, including environmental exposures, interactions with other genes, or even chance alone.

Variability in clinical and subclinical features, referred to here as phenotypic heterogeneity, is common in complex diseases and is thought to arise because of a complex genetic and environmental architecture. Such variability introduces difficulty in studying complex

disease; it is unknown if slight variations in phenotype are caused by an unknown but identifiable factor or if they carry identical risk factors and exhibit variation due to chance alone.

Environmental factors can harbor a large proportion of disease risk, as seen in many complex traits including birth defects. While environmental factors contribute to etiology, they do not completely explain the variability in complex traits, especially those with known genetic risk loci. Further exploration of the genetic variation associated with phenotypic variation, including the potential interactions between environmental and genetic factors, is of public health significance.

Additionally, in studying the variable phenotypes associated with complex traits, many distinct phenotypes are often collapsed into a broader phenotype to increase statistical power for detection of genetic loci. However, the ability to capture genetic variation responsible for subtle phenotypic variation is lost when nonhomogeneous features are misclassified as the same disease. Furthermore, in order to identify all genetic factors contributing to disease, and the mechanisms through which they interact to confer disease risk, these complex phenotypes must be studied with more granularity.

Identifying genetic sources of phenotypic variation is vital in the study of complex traits, as doing so will further the understanding of the mechanisms through which complex traits arise.

1.3 OROFACIAL CLEFTS

Cleft lip with or without cleft palate (CL/P) is a common birth defect worldwide; it is the most frequent craniofacial birth defect in humans. Approximately 1 in 800 live births has CL/P; however, the birth prevalence of CL/P varies by different ethnic groups, geographic locations and environmental exposures [64]. The highest incidence of CL/P was found in Asian and American Indian populations, followed by Caucasian populations, with African populations having the lowest incidence [76]. In developed countries, CL/P does not weigh heavily on mortality, but does result in considerable morbidity, as well as economic and

societal burden [75]. CL/P has severe consequences for affected individuals as it may inhibit or disrupt speech, facial expression, and swallowing [79].

Individuals born with CL/P may experience problems with feeding, speaking, hearing and socializing. These can be corrected to varying degrees by surgery, dental treatment, speech therapy and psychosocial intervention [17]. Despite the availability of treatments, CL/P impose a large financial and psychological affliction on affected families and society [79]. The cost per incident of CL/P is conservatively estimated to be $92,000 with a lifetime cost of treatment of $200,000, which ignores the psychosocial costs to the patients and occupational cost to parents [6] [39]. Children with CL/P also experience direct nonmedical costs, such as special education services [6]. In addition to financial costs, there are physical costs to children with CL/P; neonatal mortality is higher among children with CL/P [6].

CL/P arises when normal fetal craniofacial development fails. Cleft lip (CL) occurs when the lip fails to fuse completely in the early stages of embryogenesis. Similarly, cleft palate (CP) presents when the facial primordia, the building blocks of skulls, do not join properly. The formation of the lip is completed by the sixth week of embryogenesis, while the formation of the palate is completed by the thirteenth week [75]. A complex series of molecular events must occur for proper facial development including cell growth, migration, differentiation, and apoptosis (cell death) [39]. Similar to other congenital defects, this complex process suggests a large genetic contribution to CL/P. However, the cause of CL/P is thought to be a complex mixture of genetic predispositions and environmental exposures [64].

CL/P are considered nonsyndromic if they occur as the only abnormality; syndromic clefts are defined as those accompanied by additional structural and/or developmental irregularities [64]. In order to examine the etiology of orofacial clefts independent of other disorders, only nonsyndromic clefts are studied. It is further noted that the majority (approx. 70%) of cases of CL/P are nonsyndromic [64]. Many previous GWASs have examined the genetic role of CL/P, and many biologically plausible genes have been nominated including *IRF6*, *FGFR1*, *MAFB*, *ABCA4*, *VAX1*, Wnt signaling, *MSX1*, and *BMP* [64] [17]. However, the complex etiology of CL/P remains poorly described.

CL and CP can occur unilaterally or bilaterally, concurrently or separately. Examples of some possible types of CL/P are shown in Figure 1.1. CL and cleft lip and palate (CLP)

have historically been considered variants of the same congenital defect, differing in severity [51]. Despite the fact that CL and CLP have separate developmental geneses, they share a defect in the primary palate, motivating the combined phenotype CL/P [39].

Figure 1.1: Images of CL/P (A) bilateral cleft lip (B) unilateral cleft lip and palate (C) cleft palate

Clefts are usually regarded as simple, qualitative traits (unaffected vs. affected), although the range of physical presentations is quite large. Recently, there has been evidence suggesting that these overt clefts, in addition to subclinical phenotypes, lie on a continuum of cleft features [80]. These subclinical phenotypes may be present in unaffected relatives and would give additional genetic information about clefts overall [81]. While CL/P are visible deformations of the face, these subclinical phenotypes include lip print whorls [58], orbicularis oris (OO) muscle defects [57], and others [80]. Incorporation of these additional phenotypes may aid in explaining the complex genetic architecture of CL/P.

1.4 MOTIVATING EXAMPLES

1.4.1 CleftSeq

The CleftSeq project was the first study to perform targeted sequencing of nonsyndromic cleft lip with or without cleft palate (NSCL/P) GWAS regions. Through this we sequenced complete GWAS intervals, including non-coding and coding DNA. The 13 regions that were

sequenced were those that had been shown to be previously associated with OFCs. These 13 regions, totaling 6.3 Mb, were comprised of 9 high-priority candidates from previous GWAS and/or genome-wide linkage studies and 4 regions containing candidate genes with prior evidence of rare variants contributing to NSCL/P (Table 1.1). One thousand four hundred and ninety-eight case-parent trios from Europe, the United States, China and the Philippines were sequenced.

Because of the case-parent trio design, the transmission disequilibrium test was used to determine if there was over-transmission of risk alleles for any variant. This method is robust to population admixture. Still, we typically separate subpopulation groups for analysis, testing for association separately for Europeans and Asians, as previous studies have shown different association signals for NSCL/P in Asian and European populations. Using this method, we found three regions for functional analysis follow-up (*PAX7*, *FGFR2*, and *NOG*). We believe that this targeted sequencing of trios is powerful to identify functional variants, i.e. genetic variants which alter the function of the gene.

We also identified strong associations in multiple regions with NSCL/P in the Asian trios, but only in a single region, 8q24, in the European trios. Previous studies have shown association with many other regions in Europeans, so this study may have been underpowered to detect these. We hypothesized that many regions associated with NSCL/P are shared among different populations, and that some regions have population-specific signals. However, these hypotheses have not been rigorously tested.

Another reason that targeted sequencing was used is because it makes it possible to search for the contributions of rare variants as risk alleles for NSCL/P. However, only 2 of the 13 regions (near *NOG* and *NTN1*) showed any evidence of rare-variant over-transmission. We hypothesized that we would see many more regions with over-transmitted rare variants because of the nature of NSCL/P. Only about 50% of the heritability of NSCL/P is explained by the previously discovered genes/loci, which suggests a substantial contribution of rare variants. Notably, we did not see any rare variant signal in the four rare-variant candidate regions (*BMP4*, *FGFR2*, *MSX1*, and *PTCH1*). The rare variants in this study were analyzed with a burden-style test. This type of test cannot distinguish direction of effect or the

difference between functional and non-functional variants, which leads to decreased power to detect over-transmission in both of these situations.

Table 1.1: Overview of Regions Sequenced

	Region	Candidate Gene in Region	Target Region [GRCh37]	Size (kb)
previous GWAS hits	1p36	*PAX7*	chr1: 18,772,300 - 19,208,054	435.8
	1p22	*ARHGAP29*	chr1: 94,324,660 - 95,013,109	688.4
	1q32	*IRF6*	chr1: 209,837,199 - 210,468,406	631.2
	8q24	–	chr8: 129,295,896 - 130,354,946	1059.1
	10q25	*VAX1*	chr10: 118,421,625 - 119,167,424	745.8
	17p13	*NTN1*	chr17: 8,755,114 - 9,266,060	510.9
	17p22	*NOG*	chr17: 54,402,837 - 54,957,390	554.6
	20q12	*MAFB*	chr20: 38,902,646 - 39,614,513	711.9
previous linkage hit	9q22	*FOXE1*	chr9: 100,357,692 - 100,876,841	519.1
candidate gene regions	4p16	*MSX1*	chr4: 4,825,126 - 4,901,385	76.3
	9q22	*PTCH1*	chr9: 98,133,647 - 98,413,162	279.5
	10q26	*FGFR2*	chr10: 123,096,374 - 123,498,771	402.4
	14q22	*BMP4*	chr14: 54,382,690 - 54,445,053	62.4

We concluded that sequencing of all GWAS-implicated regions in a wide range of populations, together with functional analyses, would be necessary to fully understand the role of these genes/regions in the etiology of NSCL/P. This would give insight into shared and population-specific signals, as well as the role of rare variants in NSCL/P.

1.4.2 Multiethnic Study of Orofacial Clefts

The multiethnic OFC GWAS (also known as the Pittsburgh OroFacial Cleft [POFC] study) was a study conducted in several populations consisting of 11,727 participants recruited from 18 sites across 13 countries from North America, Central or South America, Asia, Europe, and Africa. The overall study cohort includes OFC-affected probands, their unaffected family members and controls with no known history of OFC or of other craniofacial anomalies. Thus, there are many family structures present in the OFC study, including singleton cases

and controls, sibling pairs, case-parent trios, and larger families. Currently, we have only analyzed independent (unrelated) cases, controls and trios. We conducted standard association in the cases and controls, TDT in the trios, and used inverse-variance weighted meta-analysis to estimate the effect of each variant on NSCL/P. There were a total of 6,480 participants (823 cases, 1700 controls, and 1319 case-parent trios) with European, Asian, African, and Central and South American ancestry for the aforementioned analysis. All subjects were genotyped on the same microarray with approximately 580,000 SNPs. Ideally, we would like to combine all participants into the same analysis to maximize the information available, as some signals might be lost by only examining independent trios, cases, and controls.

Again, we replicated many (but not all) previously-associated regions. Some of these regions showed evidence of shared signal between the different subpopulations (e.g. *NTN1*), but many regions appeared to be population specific. We would like to be able to quantify the heterogeneity we see in these signals between populations and assess if the difference we see is due to low power in some subpopulations with smaller sample size.

This study also collected extensive phenotypes on participants. We have detailed cleft information (type of cleft, completeness, and side affected) for each participant with a cleft. We believe that there are underlying genetic differences that contribute to these phenotypic differences we see. Very little is known about the differentiation between the cleft subtypes (isolated cleft lip, cleft lip with cleft palate, and isolated cleft palate), the side of the face affected in unilateral cleft lips, and why OFCs are more frequent in males. There is a need for a statistical test to find any underlying genetic components that contribute to these cleft differences.

1.5 SUMMARY

In this integrated dissertation we examine methods applied to genetic epidemiology studies of family-based and population-based data. This includes two main components. The first component addresses methods for combining data from mixed structure designs. The second

component examines methods of determining genetic sources of phenotypic heterogeneity in orofacial clefts.

Specifically, we address two hallmarks of the study of complex disease: (1) How can family data be combined with case-control data from the same study to improve estimates of disease association in a way that is robust to model misspecification? and (2) How can genetic sources of phenotypic heterogeneity be identified in case-control and family-based studies?

In chapter 2, we provide an overview of existing methods for analyzing mixed data types. First, methods for population-based and family-based data analysis are presented followed by two classes of methods for combining these two data types into a single analysis. Mixed data can be analyzed in two primary ways – via meta-analysis of the separate signals from the two data sources and via a joint, retrospective likelihood. An analysis of the strengths and weaknesses of these methods is given, including a practical application of one such method to a genome-wide association study for orofacial clefting.

In chapter 3, we propose an empirical Bayes-type estimator for combining mixed data structures in a retrospective likelihood to leverage the assumption of HWE among controls and parents within trios to the extent that the data supports HWE.

In chapter 4, we present an overview of phenotypic heterogeneity in the study of complex traits including methods for detecting genetic contributions to phenotypic heterogeneity.

A published study examining genetic sources of phenotypic heterogeneity in a targeted sequencing study is given in chapter 5.

In chapter 6, we apply a method for detecting genetic differences in orofacial clefting to a genome-wide meta-analysis. We present a novel approach to visually representing heterogeneity of genetic loci via the "cleft map".

2.0 GENOME-WIDE ASSOCIATION STUDIES FOR MIXED DATA STRUCTURES

2.1 INTRODUCTION

GWASs are popular tools used to detect genetic loci associated with a trait of interest. The underlying concept for GWAS is to perform a test of association for each SNP across the genome, and then examine the regions showing the most statistical significance. Thus, GWAS is applicable for use with a variety of trait distributions. Virtually any statistical model can be used to test each variant; the primary limitation is the computing power required to fit such a model for millions of genetic markers.

With dichotomous traits, the most common method for GWAS is the case-control association study which uses either a simple chi-squared test or logistic regression to examine differences in allele or genotype frequencies between cases and controls at each SNP. This type of analysis is straightforward and easy to implement, but is subject to false positive associations when there is population stratification; population allele frequency differences can be confounded with disease frequency if not properly accounted for.

A family-based design with case-parent trios uses a different statistical method to examine linkage and association of a genetic marker and the trait. The most common methods for case-parent trios include the TDT and a conditional on parental genotype (CPG) approach. Unlike the case-control approach, family-based methods are not subject to inflation of results due to population substructure, as examining transmission between parents and offspring removes any potential effect caused by population allele frequency differences. However, family-based methods have less statistical power than the population-based case-control designs for the same number of individuals studied.

More recently, methods have been developed to combine these two data structures, granting the most statistical power while still protecting against artificial inflation of results. Methods for analyzing mixed data structures leverage more information from a study with mixed data types than examining each data type separately. This chapter provides an overview of existing statistical methods for performing GWAS with cases and controls, case-parent trios, and the combination of them. We employ one such method for combining case-control and case-parent trios to explore genome-wide association in the Multiethnic OFC study. In particular, we demonstrate how utilizing mixed data structure increases the information obtained from a GWAS, without complicated methodology.

2.2 METHODS

This section first provides a basic overview of the statistical methods used for separate analysis of population-based and family-based data, then gives a survey of the methodology currently in place for combining these two data structures.

2.2.1 Case Control

The goal of a case-control study is to identify risk groups by observing outcomes; the primary interest is estimating the genotypic relative risk (GRR) given by equation 2.2.1.

$$\gamma_g = \frac{P(D=1|G=g)}{P(D=1|G=0)}, \quad g = 1, 2 \tag{2.2.1}$$

The retrospective likelihood of the observed genotypes is composed of the independent components for cases and controls where each component is a straightforward multinomial probability. The likelihood for a case-control analysis is:

$$\begin{aligned} L(\gamma_1, \gamma_2, p) &= \prod_j P(G_j|D_j=0) \prod_k P(G_k|D_k=1) \\ &= \prod_j P(G_j=g) \prod_k \frac{\gamma_g P(G_k=g)}{\sum \gamma_{g*} P(G_k=g_*)} \end{aligned} \tag{2.2.2}$$

where

$$p = \text{minor allele frequency, } P(a), \text{ in the population}$$
$$\gamma_1 = \text{relative risk of clefting for Aa compared to AA}$$
$$\gamma_2 = \text{relative risk of clefting for aa compared to AA}$$
$$D_i = \text{disease status of individual } i$$
$$G_i = \text{genotype of individual } i$$

The likelihood is parametrized by the two relative risk components and the minor allele frequency (a nuisance parameter). Typically, a genetic model is employed to reduce the GRR parameters to one parameter. The common genetic models used to do so are given in Table 2.1.

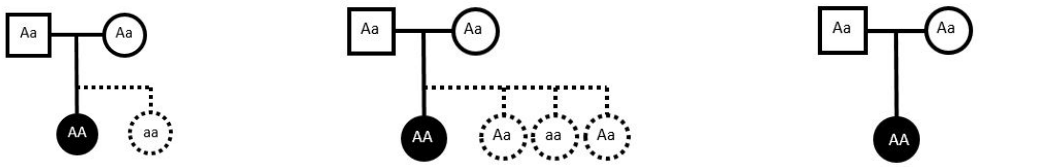
Table 2.1: Commonly used genetic models in terms of relative risk parameters

Genetic Model	γ_1	γ_2
Additive	γ	$2\gamma - 1$
Dominant	γ	γ
Recessive	1	γ
Multiplicative	γ	γ^2

2.2.2 Trios

When studying families, it is the transmission of alleles from parent to offspring that is analyzed. In this sense, parents are used as genotypic controls for their children. There are three main ways in which trios are analyzed using the idea of parents as genetic controls, each illustrated in the Figure 2.1 [74].

In the toy examples in Figure 2.1, both parents are heterozygous (Aa). (One can easily extend this to all possible parental genotypes.) One allele from each parent is transmitted to



(a) TDT (b) Conditional Approach (c) Combined Likelihood

$$L = \prod P(G_{o_i}|D_{o_i} = 1, \boldsymbol{G_{p_i}}) \quad L = \prod P(G_{o_i}|D_{o_i} = 1, \boldsymbol{G_{p_i}}) \quad L = \prod P(\boldsymbol{G_{p_i}}, G_{o_i}|D_{o_i} = 1)$$

Figure 2.1: Three main ways of estimating association in trios

the child (one A from mom, one A from dad). Example A is the TDT, which is a matched analysis comparing the child to its anti-self, the unobserved instance of a child with both non-transmitted alleles [71]. Comparing the child with its anti-self is one way to condition on the parental genotype. Similarly, Figure 2.1b shows the conditional logistic regression method which compares the proband to unobserved pseudo-siblings with all possible transmission patterns. Again, these pseudo-siblings are not observed, but the result of conditioning on the parental genotype to create a matched analysis. If an additive genetic effect is assumed, example 2.1a and example 2.1b are equivalent.

These first two methods only model the probability of the childs genotype conditional on parental genotypes, but example 2.1c models the probability of the entire trios genotype. The final method is a combined likelihood approach which jointly models the probability of parental and case genotypes. It is notable that the likelihood for example C factors into two components – one of which is the likelihood from example B.

Under the null hypotheses of all these methods, the alternative genotypes are equally likely to have been transmitted to the case; any deviation from this expected distribution in the proband is evidence of association at that locus [74].

Proceeding with the model from example C in Table 2.1, the likelihood for jointly modeling proband and parent genotypes is given by [68]

$$
\begin{aligned}
L(\gamma_1, \gamma_2, p) &= \prod_i P(G_{o_i}, \boldsymbol{G_{p_i}} | D_{o_i} = 1) \\
&= \prod_i P(G_{o_i}, \boldsymbol{G_{p_i}} | D_i = 1) P(\boldsymbol{G_{p_i}} | D_{o_i} = 1) \\
&= \prod_i \frac{\gamma_g P(G_{o_i} | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})}{\sum_{g_*} \gamma_{g_*} P(G_{o_i} = g_* | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})} \times \\
&\qquad \frac{\sum_g \gamma_g P(G_{o_i} | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})}{\sum_{g_{p*}} \sum_{g_*} \gamma_{g_*} P(G_{o_i} = g_* | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p*}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p*}})}
\end{aligned}
\tag{2.2.3}
$$

where

$$
\begin{aligned}
p &= \text{minor allele frequency, } P(a), \text{ in the population} \\
\gamma_1 &= \text{relative risk of clefting for Aa compared to AA} \\
\gamma_2 &= \text{relative risk of clefting for aa compared to AA} \\
D_{o_i} &= \text{disease status of offspring from trio } i \\
G_{o_i} &= \text{genotype of offspring from trio } i \\
\boldsymbol{G_{p_i}} = (G_{p_1}, G_{p_2}) &= \text{genotypes of parents from trio } i
\end{aligned}
$$

The likelihood is again just a multinomial likelihood for the proband, conditioning on the disease status to model GRR and parental genotypes. Using Bayes theorem and the law of total probability, this is expanded into a function of the GRR parameters and the observed genotypes. This model assumes that both alleles are equally likely to be transmitted (i.e. no meiotic drive) and that survival to birth does not depend on genotype.

2.2.3 Combined Analysis

There are two main philosophies for combining cases, controls, and trios in such a way as to model both association and transmission. The first is a meta-analysis approach, in which separate analyses are conducted for the case-control data and the trio data, and then estimates of disease risk are combined in standard meta-analysis methods. The final estimate of disease risk is a weighted combination of the individual analyses disease risks.

The other approach combines all individuals in a single likelihood, and estimates one overall disease risk. Many current methods exist for this approach, with varying assumptions and data type inclusions [56] [18]. These methods make a rare disease assumption, and furthermore assume that the disease risk is the same for probands and cases, and that all individuals are sampled from the same population. The class of likelihood-based estimators are more powerful than meta-analysis-type methods under all genetic models except dominant and whenever modeling association with rare variants [18].

2.2.3.1 Meta-Analysis Approach

Meta-analysis approaches combine distinct estimates from case-control and trio analyses. Two specific approaches are described in this section.

The method introduced by Kazeem and Farrall combines log odds ratios from the separate case-control and trio analyses into a weighted log odds ratio (ψ) [34].

$$\psi = \frac{w_{cc} log(OR_{cc}) + w_{tdt} log(OR_{tdt})}{w_{cc} + w_{tdt}} \qquad (2.2.4)$$
$$w_i = \frac{1}{Var[log(OR_i)]}$$

The corresponding test statistic is given in equation 2.2.5

$$Q = \frac{\psi^2}{Var[\psi]} \sim \chi^2 \text{ under } H_0 \qquad (2.2.5)$$

And the assessment of heterogeneity of effects is tested with equation 2.2.6.

$$X^2_H = \sum_{i=1}^{2} w_i (log(OR_i) - \psi)2 \sim \chi^2 \text{ under } H_0 \qquad (2.2.6)$$

Combining the effect estimates in this manner implicitly assumes that the effects are homogeneous. In this regard the meta-analysis approach is identical to a likelihood-based approach (discussed in the next section) which estimates only one effect, assuming that the effects are identical between cases and trio probands.

This approach is extremely easy to implement, and provides a natural interpretation of the combined odds ratio. Independent (i.e. unrelated) cases, controls, and trios are required. If there is overlap between the cases, control, and trios, the preferred method is that from Chen and Lin which uses a robust variance estimate to allow for correlated data [13].

These meta-analysis approaches are useful tools for preliminary analyses; however, they are not the most powerful methods to detect association.

The straight-forward weighted meta-analysis approach was performed with independent cases, controls, and case-parent trios from the OFC GWAS study. (Results are detailed in the section Genome-wide association study with mixed data structure for results).

2.2.3.2 Likelihood-Based Approach

Contrary to the meta-analysis approach, the likelihood approach combines cases, control, and trios into a single likelihood to obtain one estimate of disease association. The likelihood employed is a retrospective likelihood, incorporating the disease status of individuals into the probabilities within the likelihood. Using the retrospective likelihood not only accounts for the fact that cases, controls, and incident probands were recruited based on their disease status, but also establishes a framework for using genetic assumptions about the distribution of genotype probabilities to obtain more efficient estimates of the GRR parameters. The general form of the likelihood based approaches is given by [56] (2.2.7).

$$L = \prod_{i=1}^{I} P(G_{p_i}, G_i | D_{0_i} = 1) \times \prod_{j=1}^{J} P(G_j | D_j = 1) \times \prod_{k=1}^{K} P(G_k | D_k = 1) \tag{2.2.7}$$

This can be written in terms of the relative risk parameters and the minor allele frequency (as in equation 2.2.2).

$$
\begin{aligned}
L(\gamma_1, \gamma_2, p) = & \prod_{i=1}^{I} \frac{\gamma_g P(G_{o_i} | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})}{\sum_{g_*} \gamma_{g_*} P(G_{o_i} = g_* | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})} \\
& \times \frac{\sum_g \gamma_g P(G_{o_i} | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_i}})}{\sum_{g_{p_*}} \sum_{g_*} \gamma_{g_*} P(G_{o_i} = g_* | \boldsymbol{G_{p_i}} = \boldsymbol{g_{p_*}}) P(\boldsymbol{G_{p_i}} = \boldsymbol{g_{p_*}})} \\
& \times \prod_{j=1}^{J} P(G_j = g) \\
& \times \prod_{k=1}^{K} \frac{\gamma_g P(G_k = g)}{\sum_{g_*} \gamma_{g_*} P(G_k = g_*)}
\end{aligned}
\tag{2.2.8}
$$

In order to incorporate unrelated controls and unaffected parents, we make a rare-disease assumption (equation 2.2.9). Similarly we make a rare-disease approximation for cases such that equation 2.2.10 holds.

$$
P(G = g | D = 0) \approx P(G_{p,1} = g) \tag{2.2.9}
$$

$$
P(G = g | D = 1) = \frac{\gamma_g P(G_{p,1} = g)}{\sum_{g_*} \gamma_{g_*} P(G_{p,1} = g_*)} \approx \frac{\gamma_g P(G = g | D = 0)}{\sum_{g_*} \gamma_{g_*} P(G = g_* | D = 0)} \tag{2.2.10}
$$

This assumption implies that the underlying genotype probabilities are the same for the three types of data being combined, i.e. that these three samples come from the same population. Epstein et al. provide a statistical procedure for testing this assumption in the mixed data setting [18].

In addition to these assumptions, a further assumption of Hardy-Weinberg Equilibrium (HWE) can be made in order to obtain efficient GRR estimates. The assumption of HWE incorporates many assumptions including random mating, equal allele frequencies among the sexes, no mutation, no selection, etc. The statistical consequence of assuming HWE is that the genotype probabilities can be neatly defined in terms of the frequency of the major allele – $P(AA) = (1-p)^2$, $P(Aa) = p(1-p)$, and $P(aa) = p^2$, where $p = P(A)$. Using the genotype probabilities under HWE provides very efficient estimates of GRR, however, any deviation in genotypic frequencies away from HWE can cause extreme type-1 error inflation.

In order to avoid this, one possibility is to filter out variants that deviate from HWE prior to running an association test. This will theoretically improve the robustness of the procedure, but will also lead to a type-1 error inflation if the multiple tests in this two-stage procedure are not accounted for.

Another option is to forgo the assumption of HWE in the genotypic probabilities and instead model the mating type frequencies of the parents. The six possible mating-type frequencies and their corresponding frequencies, using parental mating-type estimation and under HWE, are shown in Table 2.2 [18].

Table 2.2: Genotype Frequencies

$\boldsymbol{G_p}$	G_o	under HWE	not under HWE
AA,AA	AA	$(1-p)^4$	μ_6
AA,Aa	AA	$p(1-p)^3$	$\frac{1}{2}\mu_5$
AA,Aa	Aa	$p(1-p)^3$	$\frac{1}{2}\mu_5$
Aa,Aa	AA	$p^2(1-p)^2$	$\frac{1}{4}\mu_4$
AA,Aa	Aa	$p^2(1-p)^2$	$\frac{1}{2}\mu_4$
AA,Aa	aa	$p^2(1-p)^2$	$\frac{1}{4}\mu_4$
AA,aa	Aa	$2p^2(1-p)^2$	μ_3
Aa,aa	Aa	$p^3(1-p)$	$\frac{1}{2}\mu_2$
Aa,aa	aa	$p^3(1-p)$	$\frac{1}{2}\mu_2$
aa,aa	aa	p^4	μ_1

Estimating the mating-type frequencies instead of assuming HWE provides a robust estimate of the GRR, even when HWE does not hold. However, if HWE does hold, this method loses efficiency compared to the one in which HWE is assumed.

2.3 GENOME-WIDE ASSOCIATION STUDY WITH MIXED DATA STRUCTURE

We performed the genome-wide association scan of NSCL/P for the OFC study using two subsets of our multiethnic sample and the meta-analysis approach from Kazeem and Farrall. We partitioned the total sample into two mutually exclusive analysis sets for the current study: (1) a subset of 1,319 case-parent trios (i.e. 3,957 individuals; note, from each multiplex family only one trio was chosen), and (2) a subset of 823 unrelated CL/P cases and 1,700 unrelated controls. There was no overlap between the case-parent trio group and the case-control group; the groups were considered to be independent, were analyzed separately, and then the effects were combined via meta-analysis.

The effect of each genetic variant (293,633 genotyped SNPs with MAF>1%) was analyzed within the separate groups first, then combined into a weighted effect estimate. Cases and controls were analyzed using logistic regression (including principal components of ancestry as covariates to adjust for population substructure) . The case-parent trios were analyzed with the TDT. The log odds ratios from the separate analyses were combined using inverse-variance weighting. The resulting log odds ratio was compared to a chi-square distribution with one degree of freedom, as prescribed by the Kazeem and Farrall method. We also examined the heterogeneity of the effects, and excluded the variants for which the effects were driven by one group only.

In order to detect signals common to all ancestry groups, the first scan included individuals from all populations. Then, in order to detect population-specific signals, association scans within each ancestral group (European, Asian, and Central/South American as defined by principal components of ancestry) were performed. (Stratified analysis was not performed separately in the African group due to small sample size.) The same procedure (i.e., meta-analysis of results from the trio and case-control subsets) was used for the multiethnic and population-specific scans.

Using the results from combining effects from separate case-control and TDT scans, we identified more loci than with either scan alone (Figure 2.2 & Figure 2.3 - for full results, please see the published paper [41]).

In the meta-analysis with all populations, several known NSCL/P loci reached genome-wide significance (*PAX7*, *ARHGAP29*, *IRF6*, 8q24, and *NTN1*, Figure 2.2). Only two of these regions (*IRF6* and 8q24) demonstrated genome-wide significant associations when examining the results from the separate TDT and case-control analyses.

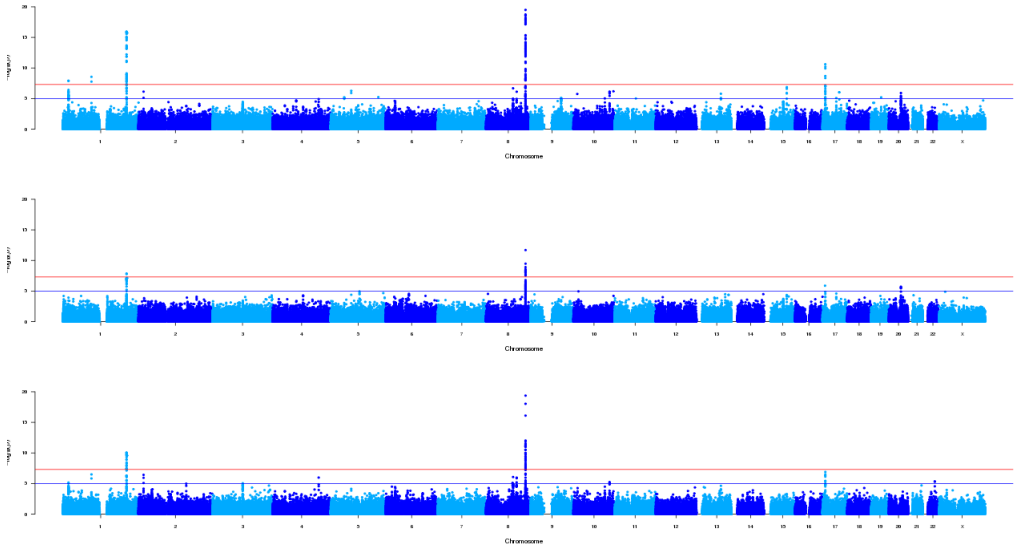


Figure 2.2: Results of the multiethnic GWAS for (A) meta-analysis, (B) TDT, (C) case-control

Furthermore, among individuals with European ancestry, we identified two genome-wide significant associations on 8q24 (a known NSCL/P locus) and 17q23a (a novel association). Three loci approached genome-wide significance: 1p36 (*PAX7*), 17p13.1 (*NTN1*), and a novel locus on 6p21.

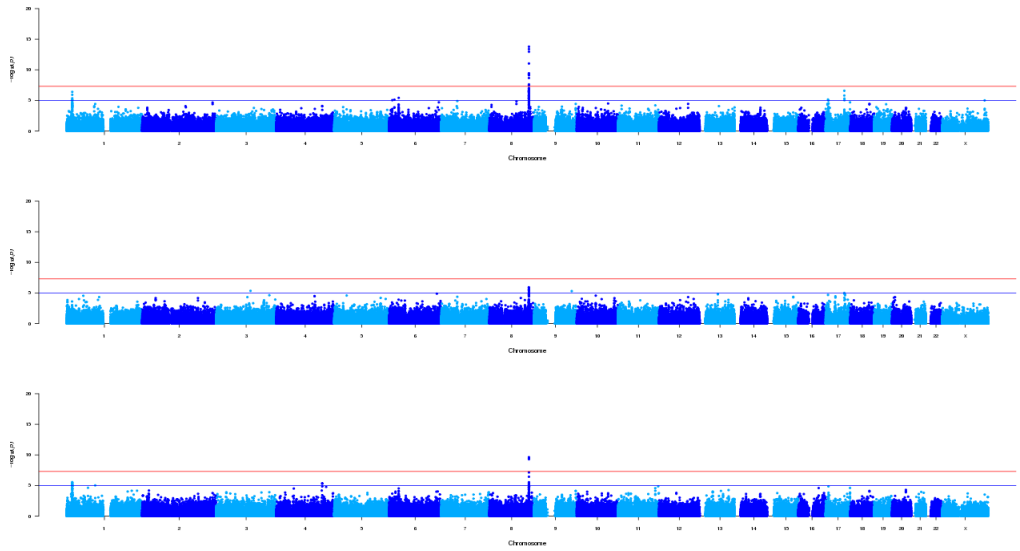


Figure 2.3: Results of the European GWAS for (A) meta-analysis, (B) TDT, (C) case-control

These association signals are summarized in Table 2.3. Regional association plots for these five significant/suggestive peaks are shown in Appendix C.

Table 2.3: Significant and suggestive loci from European GWAS

Locus	SNP	Risk Allele	TDT OR	CC OR	META OR	95% CI	P-value
1p36.13	rs9439714	C	1.62	1.38	1.52	1.30-1.89	1.91×10^{-7}
6p21.33	rs79411602	C	1.67	1.31	1.52	1.29-1.78	2.92×10^{-7}
8q24	rs72728734	G	2.22	1.84	2.04	1.70-2.44	7.33×10^{-15}
17p13.1	rs7406226	A	1.50	1.73	1.59	1.33-1.89	2.16×10^{-7}
17q23.2	rs1588366	A	1.63	2.09	1.78	1.46-2.17	1.41×10^{-8}

Analyzing the mixed data types of the OFC study yielded many genome-wide association including both known CL/P risk loci and novel loci. As demonstrated by the manhattan plots for the meta-analysis, TDT, and case-control results, the meta-analysis approach combined information from both scans resulting in strong signals where both groups demonstrated some evidence of association. The resulting effect estimates provided a natural interpretation of

the results and any heterogeneity in effects between the trio and case-control groups was detected by examining the separate effect estimates.

The OFC study utilized the meta-analysis approach because of the ease of implementation and to protect against type-1 error inflation that can arise when the strict assumptions of the likelihood-based methods are violated. Although the meta-analysis approach has reduced power compared to the likelihood-based approach, we still detected novel CL/P risk loci.

2.4 DISCUSSION

There are two broad categories of methods that incorporate data from population-based and family-based designs. The first, the meta-analysis approach, separately analyzes the information from different data structures and then uses inverse-variance-weighted meta-analysis techniques to combine those estimates. This method is easy to implement and does not require strict assumptions about the distribution of genotype probabilities. On the other hand, the likelihood-based methods require more strict assumptions about these distributions of genotype probabilities. Assuming HWE in the likelihood-based estimation techniques produces more efficient estimates of GRR than the meta-analysis approaches, but is subject to a rather dramatic increase in type-1 error. When these assumptions are violated, the estimates are inappropriately inflated. Removing this assumption of HWE in the likelihood-based methods offers one solution for balancing the trade-off between efficiency and bias.

In the example of NSCL/P, the robustness of meta-analysis approach was preferred (as demonstrated in the Multiethnic OFC GWAS) and still lead to the discovery of novel genetic loci for NSCL/P. However, if the assumptions of the likelihood-based methods could be relaxed and/or the estimates made more robust to violation of assumptions, then likelihood-based methods could be extremely useful in discovering new loci. Such methods may help elucidate the genetic architecture of clefting and other complex traits.

The methods for combining data from population-based and family-based studies discussed in this chapter are those for which transmission is directly modeled. It is worthwhile to note that many methods exist to adjust for the inclusion of related individuals, usually via a mixed model accounting for pairwise kinship of participants, but these methods do not model transmission of alleles from parent to offspring. An overview of these methods is given in [19]. Many of these methods, including one popular choice (EMMAX, [33]), are developed for quantitative phenotypes, although recently, methods have been adapted to incorporate binary traits [12]. While these methods provide useful models to account for population stratification and relatedness (including cryptic relatedness) among individuals, this dissertation is focused on methods which model transmission.

3.0 EMPIRICAL BAYES-TYPE ESTIMATION METHOD FOR MIXED DATA STRUCTURES

3.1 INTRODUCTION

A characteristic feature of GWAS is the selection of potential genetic loci to further examine after the preliminary association scan. These loci are typically selected using a p-value threshold for the association test. Choices for this threshold are determined by sample size and whether the scan is hypothesis-generating or confirmatory in nature; common thresholds are 5.00×10^{-5} and 5.00×10^{-8}. A narrow window based on genomic position, linkage disequilibrium with the most-significant SNP, and topological domains is also frequently used to select SNPs for follow-up. Lack of statistical significance beyond such thresholds in this preliminary step may exclude a positively-associated SNP from any downstream analyses. This type 2 error is particularly detrimental in hypothesis-generating scans for association as any downstream analyses, including replication efforts, will not evaluate association or biologic importance without the preliminary nomination of the variant. However, this does not grant license for high type-1 error levels. Thus, powerful methods for detecting associations that control the type-1 error rate, are necessary in order to produce potential genetic loci associated with complex traits.

Moreover, these powerful methods are needed in the mixed data structure setting. As described in chapter 2, retrospective likelihoods provide an intuitive likelihood specification for retrospective sampling including mixed data types and a framework for utilizing constraints to increase statistical power. However, existing methods for combining case-parent trios with unrelated cases and controls in a retrospective likelihood approach either assume HWE – which is efficient but biased when the data does not follow HWE – or remove any

assumption of genotype distribution – which is less efficient. The first of these methods, proposed by Nagelkerke et al. [56], made use of the HWE constraint in a retrospective likelihood combining independent cases, controls, and case-parent trios. The second method, set forth by Epstein et al. [18], removed the assumption of HWE from the estimation of GRR in a retrospective likelihood combining independent cases, controls, and case-parent trios. Without knowing the true deviation from HWE for each genetic variant, the most powerful statistical model cannot be selected *apriori*.

An ideal estimator would be a combination of these two methods such that it utilized a HWE equilibrium assumption to shrink estimates but only to the extent that the data supported the HWE assumption. The estimate would not be prone to the bias and corresponding type-1 error inflation from a constraint of HWE but would not lose efficiency in estimating the GRR by estimating many nuisance parameters. We propose such an estimator – an empirical Bayes-type shrinkage estimator – which combines the constrained and unconstrained estimation approaches previously described.

The proposed estimator maximizes statistical power and avoids increased type-1 error through the use of a data-adaptive method leveraging the HWE assumption. It achieves this by estimating the GRR parameter via "shrinkage" of the model-free estimator (not assuming HWE) towards a model-based estimator (assuming HWE). The procedure described here was introduced by Mukherjee and Chaterjee and applied to retrospective case-control studies by Luo et al. [54] [48]. We extend this method to estimate genetic risk incorporating case-parent trios simultaneously with independent cases and controls. A key feature of the proposed estimator is that it relaxes the model constraints through a completely data-adaptive shrinkage estimation approach, which controls the number of false positives due to departure from HWE.

We evaluate the performance of the proposed method compared with a constrained and unconstrained method using both simulated genetic data and real data from the Multiethnic OFC study. In particular, the application of this method to the Multiethnic OFC study provides insight into the performance of the proposed estimator on a genome-wide scale.

3.2 METHODS

3.2.1 Assumptions and Notation

We will assume a sample a cases, controls, and case-parent trios that are all genotyped at a SNP. For generalization, we denote the two alleles at the SNP A and a. G represents the observed genotype of the unrelated individuals (i.e. cases and controls), G_o the observed genotype of the affected offspring of the trio, and $\boldsymbol{G_p} = (G_{p1}, G_{p2})$ the unordered genotypes of the parents of the trio. Each genotype, g, is coded as the number of copies of the minor allele, a, taking values 0, 1, and 2. We will further assume no Mendelian errors in the trios and that no parent within a trio is affected. We denote affection with D, equaling 1 for affected and 0 for unaffected individuals. We will again use a rare-disease approximation as in equation 2.2.9 to model genotype probabilities.

3.2.2 Likelihood Formation

Consider again the the retrospective likelihood for combining independent cases, control, and case-parent trios (2.2.7).

This likelihood depends on the GRR, γ, and the genotype probabilities from controls and trio parents. We use a reparameterization to define the genotype frequencies (p_0, p_1, and p_2) in terms of HWE parameters θ and ω (3.2.1) [45].

$$\begin{aligned} \theta &= \frac{1}{2} log(\frac{4p_0p_2}{p_1^2}) \\ \omega &= \frac{1}{2} log(\frac{p_0}{p_2}) \end{aligned} \tag{3.2.1}$$

This defines the genotype probabilities of controls in terms of their HWE parameters θ and ω (3.2.2).

$$
\begin{aligned}
p_{00} = P(AA) &= \frac{e^{2\omega}}{1 + e^{2\omega} + 2e^{\omega-\theta}} \\
p_{01} = P(Aa) &= \frac{2e^{\omega-\theta}}{1 + e^{2\omega} + 2e^{\omega-\theta}} \\
p_{02} = P(aa) &= \frac{1}{1 + e^{2\omega} + 2e^{\omega-\theta}}
\end{aligned} \tag{3.2.2}
$$

The deviation from HWE is measured with the parameter θ. Values close to zero indicate mild to no deviation from HWE, whereas larger absolute values indicate violation of the HWE assumption. Specifically, $\theta > 0$ corresponds to excess homozygosity whereas $\theta < 0$ corresponds to excess heterozygosity.

We use the genotype frequencies from parents of case probands and unrelated controls to estimate the HWE parameters in two ways – unconstrained (3.2.3) and constrained under HWE (3.2.4).

$$
\begin{aligned}
\hat{\omega} &= \frac{1}{2} log(\frac{n_0}{n_2}) \\
\hat{\theta} &= \frac{1}{2} log(\frac{4n_0 n_2}{n_1^2})
\end{aligned} \tag{3.2.3}
$$

$$
\begin{aligned}
\tilde{\omega} &= log(\frac{2n_0 + n_1}{n_1 + 2n_2}) \\
\tilde{\theta} &= 0
\end{aligned} \tag{3.2.4}
$$

Given θ and ω, we characterize the genotype frequency of cases and trios in terms of genotype frequencies of controls and parents (Table 3.1, Table 3.2, and Table 3.3).

Thus, the likelihood for jointly modeling the GRR for cases, controls, and case-parent trios $L = L(\boldsymbol{\beta}, \theta, \omega)$ is a function of the relative risk and HWE parameters, where β is the $log(GRR)$.

To model GRR parameters, we could consider an unstructured model that allows for estimation of γ_1 and γ_2 without the assumption of a genetic model. However, assuming an additive genetic model reduces the number of parameters to estimate and eases computation.

Table 3.1: Genotype frequencies of controls.

genotype	count	$P(G_{ctrl}\|D=0)$
0	N_{00}	p_{00}
1	N_{01}	p_{01}
2	N_{02}	p_{02}

Table 3.2: Genotype frequencies of cases.

genotype	count	$P(G_{case}\|D=1)$
0	N_{10}	$\frac{p_{00}}{p_{00}+\gamma_1 p_{01}+\gamma_2 p_{02}}$
1	N_{11}	$\frac{p_{01}}{p_{00}+\gamma_1 p_{01}+\gamma_2 p_{02}}$
2	N_{12}	$\frac{p_{02}}{p_{00}+\gamma_1 p_{01}+\gamma_2 p_{02}}$

The additive model (i.e. $\gamma_1 = \gamma$, $\gamma_1 = 2\gamma - 1$) is assumed here to reduce labor of computation and because of its widespread use in association studies of orofacial clefting, our primary application of this method.

Let $\hat{\beta}(\theta)$ denote the maximum likelihood estimate of β for a fixed θ, and $\hat{\beta}^0(\theta = 0)$ denote the maximum likelihood estimate of β subject to the constraint that $\theta = 0$ (i.e. HWE holds for controls and parents). Both of these estimates can be obtained through standard maximum likelihood procedures, although it is worth noting that the estimates are obtained through iteratively through numerical optimization techniques as the formula for the MLEs cannot be expressed in closed form.

3.2.3 Construction of the Empirical Bayes-Type Estimator

We propose to combine $\hat{\beta}$ and $\hat{\beta}^0$, the constrained and unconstrained estimators, using an empirical Bayes-type shrinkage estimation approach as in [54].

Table 3.3: Genotype frequencies of trios.

parental genotype	offspring genotype	count	$P(\boldsymbol{G_{parents}}, G_{offspring} \mid D_{offspring} = 0)$
0,0	0	N_{000}	p_{00}^2/R
0,1	0	N_{010}	$p_{00}p_{01}/2R$
0,1	1	N_{011}	$\gamma_1 p_{00}p_{01}/2R$
1,1	0	N_{110}	$p_{01}^2/4R$
1,1	1	N_{111}	$\gamma_1 p_{01}^2/2R$
1,1	2	N_{112}	$\gamma_2 p_{01}^2/4R$
0,2	1	N_{021}	$\gamma_1 p_{00}p_{02}/R$
1,2	1	N_{121}	$\gamma_1 p_{01}p_{02}/2R$
1,2	2	N_{122}	$\gamma_1 p_{01}p_{02}/2R$
2,2	2	N_{222}	$\gamma_2 p_{02}^2/R$
$R = p_{00}^2 + \frac{1}{2}p_{00}p_{01} + \frac{1}{4}p_{01}^2 + \gamma_1(\frac{1}{2}p_{00}p_{01} + \frac{1}{2}p_{01}^2 + p_{00}p_{02} + \frac{1}{2}p_{01}p_{02}) + \gamma_2(\frac{1}{4}p_{01}^2 + \frac{1}{2}p_{01}p_{02} + p_{02}^2)$			

In order to construct the empirical Bayes-type estimator we assume an underlying distribution for a hyperparameter, called θ, with expectation zero and some variance, τ^2. Thus, the conditional distribution of $\hat{\theta}|\theta$ has the same distribution as θ, with mean θ and variance σ_θ^2. By the rules of conditional expectation and variance, $\hat{\theta}|\theta$ has mean zero and variance $\tau^2 + \sigma_\theta^2$.

Importantly, only the hyperparameter is assumed to have an underlying distribution, and other parameters are estimated with standard maximum likelihood methods, granting the name Bayes-type estimation. The general formulation of a Bayes-type estimator is a weighted average of a constrained estimate, $\hat{\beta}^0$, and an unconstrained estimate, $\hat{\beta}$ (3.2.5).

$$\hat{\beta}_{EB} = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\tau}^2 + \hat{\sigma}_\theta^2}\right)\hat{\beta}^0 + \left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_\theta^2}\right)\hat{\beta} \tag{3.2.5}$$

Considering a general empirical Bayes-type estimator as some function $\phi = f(\theta)$, where θ is the nuisance parameter (here, the HWE parameter) is assumed to have a Normal dis-

tribution with some specific variance-covariance matrix, $\boldsymbol{A}$. Applying Taylor's expansion of ψ about $\theta = 0$, the prior on ψ can be approximated with a Normal distribution with mean $f(0)$ and variance-covariance matrix $V_\phi = f'(0)^T \boldsymbol{A} f'(0)$ [54]. Then the formulation of the empirical Bayes-type estimator ϕ is given in 3.2.6.

$$\hat{\phi} = f'(0)^T \boldsymbol{A} f'(0)[\hat{V}_\phi + f'(0)^T \boldsymbol{A} f'(0)]^{-1} f(\hat{\theta}) + \hat{V}_\phi[\hat{V}_\phi + f'(0)^T \boldsymbol{A} f'(0)]^{-1} f(0) \tag{3.2.6}$$

Thus, using 3.2.6, the formulation of the empirical Bayes-type estimator for combining the constrained GRR estimate [under HWE, $\hat{\beta}^0(\theta = 0)$], with the unconstrained GRR estimate $[\hat{\beta}(\hat{\theta})]$ is given in equation 3.2.7. Mukherjee and Chatterjee provide a detailed rationale and general formulation for empirical Bayes-types estimators [54].

$$\hat{\beta}_{EB} = \hat{\beta} - \hat{V}_{\hat{\beta}}(\hat{V}_{\hat{\beta}} + \hat{\theta}^2 \hat{\Delta}^T \hat{\Delta})^{-1}(\hat{\beta} - \hat{\beta}^0) \tag{3.2.7}$$

where $\hat{V}_{\hat{\beta}}$ is the estimated variance-covariance matrix of $\hat{\beta}$ and $\hat{\Delta} = \frac{\delta \hat{\beta}(\theta)}{\delta \theta}|_{\theta=0}$.

This proposed estimate is a weighted combination of the constrained and unconstrained estimates. This new estimate should have the ideal properties of being closer to the constrained estimate when HWE is indeed true in the population, and closer to the unconstrained estimate otherwise.

Intuitively the gradient function, $\hat{\Delta}$, represents rate of change of the unconstrained estimator in direction of θ at the point when the genotype frequencies are under HWE. The influence that $\hat{\theta}$ has in the weighting of the empirical Bayes-type estimator depends on $\hat{\Delta}$, such that more severe deviation from HWE weights the estimator more heavily towards the unconstrained estimate and vice versa. We employ a first-order Taylor expansion to approximate the $\hat{\Delta}$ (3.2.8).

$$\hat{\Delta} \approx \frac{1}{\hat{\theta}} \left(\hat{\beta} - \hat{\beta}^0 \right) \tag{3.2.8}$$

In order to obtain the asymptotic properties of $\hat{\beta}_{EB}$, we note that the score function, that is the derivative of the log-likelihood with respect to the GRR, can be expressed as in equation 3.2.9.

$$\sum_{i \,\epsilon \text{ cases, probands}} \left[\left(\frac{x_i N_{1i}}{p_{1i}} \right) \left(\frac{\delta p_{1i}}{\delta \beta} \right) + \left(\frac{(1-x_i) N_{pi1,pi2,i}}{p_{pi1,pi2,i}} \right) \left(\frac{\delta p_{gi1,gi2,i}}{\delta \beta} \right) \right] = 0 \tag{3.2.9}$$

where $x_i = I(i \,\epsilon$ cases$)$

$1 - x_i = I(i \,\epsilon$ probands$)$

$N_{1i} =$ number of cases of genotype i

$N_{pi1,pi2,i} =$ number of trios with parent genotypes $pi1$ and $pi2$, and proband genotype i

Consequently, we can construct a partial M-estimator of the form $\sum \boldsymbol{\psi} = 0$ using equation 3.2.9 and creating an extension for the empirical Bayes-type estimator (3.2.10). In this formulation, we are ignoring variation in $\hat{V}_{\hat{\beta}}$ and $\hat{\Delta}$ and treating them as known. Logically, the variance of these quantities approaches zero as the sample size increases, so this assumption that they are fixed is only inappropriate in small sample sizes, which is uncommon for genome-wide association studies in general.

$$\sum_{i \,\epsilon \text{ cases, probands}} \begin{bmatrix} \left(\frac{x_i N_{1i}}{\tilde{p}_{1i}(\beta^0)} \right) \left(\frac{\delta \tilde{p}_{1i}(\beta^0)}{\delta \beta^0} \right) + \left(\frac{(1-x_i) N_{pi1,pi2,i}}{\tilde{p}_{gi1,gi2,i}(\beta^0)} \right) \left(\frac{\delta \tilde{p}_{gi1,gi2,i}(\beta^0)}{\delta \beta^0} \right) \\ \left(\frac{x_i N_{1i}}{p_{1i}(\beta)} \right) \left(\frac{\delta p_{1i}(\beta)}{\delta \beta} \right) + \left(\frac{(1-x_i) N_{pi1,pi2,i}}{p_{gi1,gi2,i}(\beta)} \right) \left(\frac{\delta p_{gi1,gi2,i}(\beta)}{\delta \beta} \right) \\ \beta - \hat{V}_{\hat{\beta}} (\hat{V}_{\hat{\beta}} + \hat{\theta}^2 \hat{\Delta}^T \hat{\Delta})^{-1} (\beta - \beta^0) - \beta_{EB} \end{bmatrix} = 0 \tag{3.2.10}$$

where $\tilde{p}(\beta^0) =$ the genotype frequency under the constrained model

$\tilde{p}(\beta) =$ the genotype frequency under the unconstrained model

Thus, a straightforward application of M-estimation theory provides that equation 3.2.11 holds [72].

$$\sqrt{N}(\hat{\beta}_{EB} - \beta_{EB}) \overset{o}{\sim} N(0, V_N(\beta_{EB})) \tag{3.2.11}$$

Using this theory we can construct a robust sandwich estimate of the variance through equation 3.2.12.

$$V_N(\beta_{EB}) = A_N^{-1} B_N [A_N^{-1}]^T \tag{3.2.12}$$

$$\text{where } A_N = \frac{1}{N} \sum_{1}^{N} -\boldsymbol{\psi'}$$

$$B_N = \frac{1}{N} \sum_{1}^{N} \boldsymbol{\psi}\boldsymbol{\psi}^{\boldsymbol{T}}$$

The construction of asymptotic variance-covariance matrix is given in Appendix B.

We can then construct a simple Wald-type test using the resulting estimate of $\hat{\beta}_{EB}$ and $\hat{V}(\hat{\beta}_{EB})$ via equation 3.2.13 to test the null hypothesis of no association, $H_0 : \beta_{EB} = 0$. Given the theory from M-estimation, this test statistic follows a χ^2 distribution with one degree of freedom.

$$Q = \frac{\hat{\beta}_{EB}^2}{\hat{V}(\hat{\beta}_{EB})^2} \tag{3.2.13}$$

3.3 RESULTS

3.3.1 Simulation Study

We perform simulations of the empirical Bayes-type estimator, compared to the constrained and unconstrained estimators, using the robust variance estimates from the M-estimation framework. Under the null hypothesis (i.e. $\gamma = 0$), we simulated 10,000 genetic variants for equal sample sizes ($N_{cases} = 500$, $N_{controls} = 500$, $N_{trios} = 500$), varying θ ($\theta = 0$, $0.5log(1.2)$, $0.5log(1.6)$, and $0.5log(2.0)$ representing no, small, modest, and large deviations

from HWE) and minor allele frequency ($MAF = 0.1, 0.2, 0.3$). For each of these settings, the robust constrained, unconstrained, and empirical Bayes-type estimators are extremely conservative (average results shown in Table 3.4). Under an alternative hypothesis of modest genetic association (i.e. $\gamma = 1.5$), we simulated the same sample size, HWE parameter, and minor allele frequency settings, and observed the same extreme conservativeness in test of all three robust estimators (Table 3.5). In each simulation, under the null and alternative hypotheses, the empirical Bayes-type estimates of the GRR behave as one would expect: they are equivalent to the constrained estimate of GRR when $\theta = 0$ and move closer to the unconstrained estimate as θ increases. However, for large numbers of genetic variants and/or large sample sizes, the computational time for calculating the robust variance estimates for the constrained, unconstrained, and empirical Bayes-type estimators was markedly increased compared to standard genome-wide methods.

Table 3.4: Average simulation results under null hypothesis, $\gamma = 0$.

θ	MAF	$\hat{\beta}^0$	$\hat{V}(\hat{\beta}^0)$	$\hat{\beta}$	$\hat{V}(\hat{\beta})$	$\hat{\beta}_{EB}$	$\hat{V}(\hat{\beta}_{EB})$
0	0.1	0.2379	80.9960	0.2378	81.0672	0.2379	81.0671
0	0.2	0.1893	20.9505	0.1892	20.9582	0.1892	20.9573
0	0.3	0.1460	10.1467	0.1458	10.1466	0.1460	10.1468
$0.5log(1.2)$	0.1	0.2360	80.8334	0.2367	81.6750	0.2360	81.6679
$0.5log(1.2)$	0.2	0.1879	20.8835	0.1903	21.2062	0.1879	21.1972
$0.5log(1.2)$	0.3	0.1454	10.1245	0.1498	10.2917	0.1454	10.2835
$0.5log(1.6)$	0.1	0.2356	80.7066	0.2377	83.0826	0.2356	83.0619
$0.5log(1.6)$	0.2	0.1855	20.6941	0.1922	21.6117	0.1855	21.5865
$0.5log(1.6)$	0.3	0.1420	9.9943	0.1537	10.4574	0.1420	10.4351
$0.5log(2.0)$	0.1	0.2331	80.2845	0.2364	84.1715	0.2331	84.1381
$0.5log(2.0)$	0.2	0.1838	20.6200	0.1940	22.0719	0.1838	22.0317
$0.5log(2.0)$	0.3	0.1405	9.9439	0.1581	10.6808	0.1405	10.6461

Table 3.5: Average simulation results under alternative hypothesis, $\gamma = 1.5$.

θ	MAF	$\hat{\beta}^0$	$\hat{V}(\hat{\beta}^0)$	$\hat{\beta}$	$\hat{V}(\hat{\beta})$	$\hat{\beta}_{EB}$	$\hat{V}(\hat{\beta}_{EB})$
0	0.1	0.5438	208.9158	0.5438	209.2517	0.5438	209.2515
0	0.2	0.4964	61.1924	0.4963	61.2595	0.4963	61.2572
0	0.3	0.4529	33.3825	0.4526	33.3816	0.4529	33.3823
$0.5log(1.2)$	0.1	0.5410	208.7897	0.5417	212.3202	0.5410	212.3079
$0.5log(1.2)$	0.2	0.4946	60.0873	0.4971	61.6559	0.4945	61.6400
$0.5log(1.2)$	0.3	0.4514	33.0785	0.4560	34.1232	0.4514	34.1002
$0.5log(1.6)$	0.1	0.5412	207.1902	0.5432	217.0578	0.5412	217.0224
$0.5log(1.6)$	0.2	0.4912	59.2110	0.4982	63.8355	0.4912	63.7881
$0.5log(1.6)$	0.3	0.4463	31.9479	0.4589	34.8002	0.4463	34.7376
$0.5log(2.0)$	0.1	0.5371	205.0775	0.5405	221.7345	0.5371	221.6768
$0.5log(2.0)$	0.2	0.4885	58.1676	0.4993	65.2815	0.4885	65.1995
$0.5log(2.0)$	0.3	0.4437	31.4537	0.4626	36.0093	0.4437	35.9122

3.3.2 Application to Genome-wide Study of Orofacial Clefts

We applied this method to a sample of 170 cases, 835 controls, and 1050 individuals from case-parent trios (i.e. 350 trios) of European decent (as identified through principal components of ancestry) from the Multiethnic OFC study. Association at each of 258,543 genotyped SNPs with MAF $> 5\%$ was examined using three methods: (1) the constrained estimation approach using the robust sandwich variance estimate, (2) the unconstrained estimation approach using the robust sandwich variance estimate, and (3) the empirical Bayes-type estimation approach using the robust sandwich variance estimate.

None of these methods demonstrates any statistical significance (i.e. $p > 0.05$ for all variants). However, if the ranked order of the variants is considered rather than a p-value threshold, the highest ranked variants are those from the regions which demonstrated genome-wide

statistical significance (i.e. 8q24 and 17p13.1) in the preliminary association scan in chapter 2 as shown in Table 3.6.

Table 3.6: Top 20 variants from empirical Bayes-type estimation.

Ranking	SNP	CHR	BP	P-value (European OFC meta-analysis)
1	rs7018093	8	129891232	1.40×10^{-6}
2	rs1850889	8	129890405	1.45×10^{-6}
3	rs7841974	8	129888565	1.86×10^{-6}
4	rs2056314	8	129875260	2.00×10^{-6}
5	rs7010446	8	129874453	2.06×10^{-6}
6	rs756122	8	129912740	2.64×10^{-6}
7	rs2395865	8	129903689	3.88×10^{-6}
8	rs10100830	8	129893934	2.98×10^{-6}
9	rs2119756	8	129898369	4.54×10^{-6}
10	rs1519851	8	129895819	4.66×10^{-6}
11	rs2395864	8	129903563	4.91×10^{-6}
12	rs4733659	8	129910410	6.06×10^{-6}
13	rs4733532	8	129881299	5.14×10^{-6}
14	rs1519849	8	129896967	9.51×10^{-6}
15	rs9297779	8	129986237	4.02×10^{-5}
16	rs7844704	8	129845635	1.64×10^{-5}
17	rs1519850	8	129896821	4.87×10^{-5}
18	rs3760257	17	61496471	9.25×10^{-6}
19	rs6470670	8	129913448	5.89×10^{-5}
20	rs873761	8	129863533	3.95×10^{-5}

3.4 DISCUSSION

We constructed an empirical Bayes-type estimator as a weighted combination of a constrained estimate (under HWE) and unconstrained estimate. The resulting estimate cannot be solved exactly and linear approximations must be employed. In order to obtain the distribution of the resulting estimate, a partial M-estimator was constructed and the robust sandwich variance estimate calculated.

Because of the need for robust variance estimates of the empirical Bayes-type estimator, the efficiency gained by employing a shrinkage-estimate is lost. This is seen in the all three estimators – constrained, unconstrained, and empirical Bayes-type – using the robust variance estimates. Each method is extremely conservative, and only provides a ranked order of the variants.

These results are contrary to the presentation of empirical Bayes-type estimators to improve efficiency [54] [48]. We believe this is largely due to the robust variance estimate, which is indeed robust against model misspecification, but may not be useful in the context of genome-wide analyses due to the inefficiency of the estimates and the computational burden. Furthermore, the decrease in efficiency may also be due to the multiple approximations that are required to formulate the empirical Bayes-type estimator (e.g. the approximation of $\hat{\Delta}$) 3.2.8. While these approximations behave well in neighborhoods of the point of expansion, they may not achieve the same properties when used to construct an estimator which will be tested in a genome-wide setting.

Despite these concerns of efficiency, the application to the Multiethnic OFC GWAS demonstrates that the statistical ranking of the variants was preserved in general. Thus, the estimator is in fact testing for genetic association and does identify regions with strong, known effects, albeit with much less efficiency than estimators with non-robust variance estimates.

An additional possibility for testing association with an empirical Bayes-type estimation procedure without the robust variance estimate from M-estimation framework, would be to perform permutations at each variant, comparing the estimate to it's empirical distribu-

tion. While this would provide a smaller variance estimate, it would further increase the computational burden.

Given these considerations, other existing methods may be more useful to analyze genome-wide associations from mixed data types. As demonstrated in chapter 2, the meta-analysis approach combining effect estimates from logistic regression and TDT is easy to implement and useful. If a retrospective likelihood is desired to combine cases, controls, and trios, the SCOUT software which implements the likelihood-based method without the assumption of HWE is preferred over methods assuming HWE [18].

4.0 GENETICS OF PHENOTYPIC HETEROGENEITY

In the study of genetic association, a distinction is made between simple and complex traits. Simple traits are those more-or-less following a Mendelian inheritance pattern, and are usually controlled by a single genetic factor with strong effect. Conversely, complex traits often exhibit a sporadic inheritance pattern and often involve many genetic and environmental factors of more modest effect sizes. Common traits are usually thought to have complex etiology with many genes affecting the trait. Identifying these genetic variants associated with common traits depends on the statistical power to detect them in association studies. While variants showing strong genetic effects are easily detected, it requires large sample sizes to detect modest effects. Furthermore, higher-frequency variants have increased power of detection than lower-frequency variants with the same effect size. Thus, preliminary studies of complex traits mostly identified associations with common variants [84]. This leads to the common disease-common variant (CDCV) hypothesis that has been widely used to study complex traits. The hypothesis specifies that even though common diseases/traits are usually determined by many genetic loci, most of the genetic risk is attributable to common variants, and each genetic locus typically has one common variant [84]. This implies that common genetic variants, which are more readily detectable than rare variants, are largely responsible for variation in complex diseases and traits. However, common variants have increased statistical power for detecting associations; the contributions of rare variants to common, complex disease may just be understudied and underpowered. While many common-variant associations for complex traits have been detected, common variation still has not accounted for all of the heritability present in many complex traits. The CDCV hypothesis may not be capturing the underlying genetic etiology of complex traits, which

could involve rare variant contribution. Further discussion of the CDCV hypothesis is given by Pritchard and Cox [59].

Beyond the frequency of variants associated with complex traits, the heterogeneity of the trait itself may influence the ability to identify genetic associations. The genetic architecture of complex traits can involve numerous genetic loci (with rare or common variants) affecting many biologic processes. Typically, complex traits exhibit more heterogeneous phenotypes which can introduce noise into association analyses if the phenotypic variation is not entirely related to genetic variation. Ideally, more homogeneous groups (based on genetic etiology) would be constructed to study only the phenotypic variability associated with genetic variability. However, these groups are nearly impossible to construct without *a priori* knowledge of the causal genetic mechanisms.

Statistically, the problem of phenotypic heterogeneity has been addressed in numerous ways using classic statistical models. However, these methods have not been traditionally employed to discover genetic differences in subphenotypes. In the following chapter, an overview of the statistical methods capable of addressing phenotypic heterogeneity is given, along with a comparison of how these methods operate to detect genetic sources of phenotypic heterogeneity.

4.1 INTRODUCTION

A hallmark of common, complex disease is phenotypic heterogeneity, which arises when one disease or disorder presents itself in multiple different ways. These differences can be based on a number of factors including severity of disease, age of onset, and presence of disease subtypes. The sources of this variability in phenotype are often unknown: they may be due to genetic, environmental, or unknown variation. For complex diseases with known genetic factors, it is very plausible that genetics also play a role in phenotypic variability.

To investigate this phenotypic heterogeneity, subphenotype groups are often defined. These groups can be defined by categorizing a disease by severity, serological thresholds, or even more overt subtypes that exist. Categorizing a phenotype into subphenotypes and

studying the genetic underpinnings of each subgroup will provide meaningful insight into the genetic architecture of a complex trait. This is especially the case when subphenotypes have distinct causal mechanisms – only through comparing the subphenotypes would these mechanisms be discovered. The study of subphenotypes will help elucidate the biologic mechanisms operating to influence the trait. For example, nonsyndromic cleft lip with or without cleft palate is a highly heterogeneous trait. It is hypothesized that some types of clefts may share genetic etiology while others may have unique causes. Additionally, it may be the accumulation of genetic risk factors that cause a cleft or are responsible for cleft type differentiation. The biologic model that these risk factors operate within cannot be identified without first understanding genetic sources of phenotypic heterogeneity.

Subgroup analysis essentially tests for three possible genetic sources of phenotypic variation, described in Figure 4.1. The first, referred to as shared genetic variants, affects subphenotypes in the same manner, i.e. when they share a genetic etiology. The second, referred to as subtype-specific genetic variants, increases disease risk or susceptibility in one specific subphenotype group while the other caries the same baseline risk as controls. The last possibility considered, referred to as modifier genetic variants, confers significantly different risk between subphenotypes. Modifiers can work in one of two primary ways: (1) increasing disease susceptibility overall, but with stronger effect in one subgroup - referred to as "gradient"; or (2) not changing disease susceptibility overtly, but increasing risk in one subtype while decreasing risk for another - referred to as "opposite".

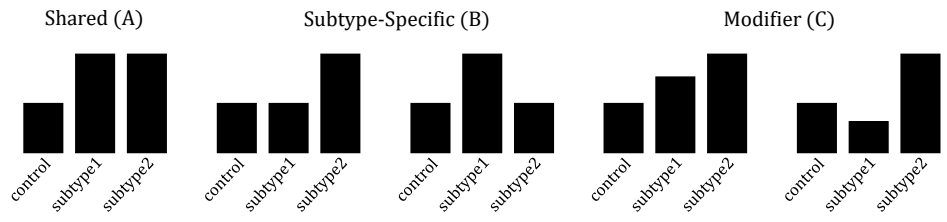


Figure 4.1: Allele frequencies for possible genetic sources of phenotypic heterogeneity: (A) Shared, (B) Subtype-Specific, and (C) Modifier.

Importantly, each of these models are for one genetic locus at a time. Thus, a disease may have multiple associated loci working in concert to affect phenotypic variation through individual contributions to overt disease risk, subtype-specific risk, and/or subphenotype differences. Detecting these different types of heterogeneity necessitates multiple methods, as many methods are only powered to detect one type of heterogeneity.

4.2 METHODS

There are numerous methods to examine the genetic sources of phenotypic heterogeneity and each method addresses a very specific hypothesis and study design. Some methods require a case-control study design while a few other lend themselves to family-based or mixed study designs. Furthermore, many methods require genotype-level data – extracting inference from association methods – while others used summary statistics post hoc. In general, there is a lack of consensus for the appropriate method to use given a hypothesis of phenotypic heterogeneity. Each hypothesis for phenotypic heterogeneity (see Figure 4.1) requires a unique contrast to test for that type of genetic variation. The following section addresses this gap in knowledge by comparing the existing methods for testing phenotypic heterogeneity and the corresponding philosophical question they address.

4.2.1 Genotype-level Tests

The primary types of methods considered here are (1) a pooled approach which combines subphenotypes for analysis, (2) a separate approach which analyzes subphenotypes independently, (3) a case-only approach which directly compares one subtype to another, (4) a likelihood approach for genome-wide scans of multiple types of heterogeneity, and (5) a gene-by-environment framework that leverages information from case-parent trio designs. Each of these methods requires genotype-level data.

4.2.1.1 Pooled Method

Combining subphenotypes into a broader phenotypic definition is a common approach to increase statistical power to detect association. However, it is not a method for exploring phenotypic heterogeneity; rather, it models perfect homogeneity of effects. Nevertheless, employing the pooled method is useful in examining genotype-phenotype association. The association signal detected with the pooled method may be caused by a shared variant signal, by a subtype-specific signal, or by a type 1 modifier signal (increasing risk overall but with different risk conferred between subphenotypes).

When it is hypothesized that subtypes share genetic etiology, combining similar phenotypes for analysis is common. If subphenotypes have the same genetic underpinnings, then pooling these cases yields the most power to detect genetic variation that is associated with this pooled phenotype. If the true genetic signal is subtype-specific, it may be possible to detect the signal in a pooled analysis but the effect will be watered down by the presence of the unassociated second subphenotype. Finally, a modifier genetic effect could be detected if the variant conferred at least some disease risk to each subphenotype, but the disease risk is assumed to be identical between subphenotypes and is thus biased. If the modifier results in effects with opposite direction, this method will fail to detect any association.

As previously mentioned, this type of analysis is commonly used for complex traits due to limited sample sizes when phenotypes are broken down into more homogeneous groups. While association signals using a pooled approach can be driven by a shared genetic etiology, it is not reasonable to assume this is true without follow-up. This is particularly problematic when one subphenotype occurs more frequently than another. As a result of disparate sample sizes, little information is obtained from the less-frequent subphenotype and results are driven by the more frequent one. In any case, further steps should be taken to narrow down the source of the effect – whether it is shared, driven by one subphenotype, or different between subphenotypes.

4.2.1.2 Separating Method

Identifying sources of unique pathophysiology is the goal of methods which separate subphenotypes into distinct groups for different analyses. In this approach, only genetic markers that have a signal in the more homogeneous phenotype groups will be detected.

When the genetic markers are truly associated with both subphenotypes, the separating approach will only detect association when the sample sizes are sufficient. In general, it is less powered to detect shared effects than the pooled approach. Thus, only when the association signal is strong (i.e., large effect) or the sample sizes very large, will the separate approach detect any association shared between both subphenotypes. In the case when the increased disease risk is unique to one subphenotype, the subphenotypes should be analyzed separately to capture the distinction. Both modifier variants with varying degrees of increased risk in both subphenotypes and those with opposite effects may be detected using this method, although there would be no indication of the modifying nature of the locus.

A natural next step for this type of analysis is to compare the results from each of the separate analyses. However, direct comparison of the resulting p-values from separate analyses is not a valid method for examining heterogeneity, as p-values are dependent on sample size and phenotypic distribution. Furthermore, qualitative comparison of p-values and effect estimates does not give any information about potential differences between the subphenotype specific analyses. Methods for comparing these results are discussed in the summary-level tests section.

Similar to the pooled approach, it is feasible to perform separate analyses for subphenotypes with virtually any study design and statistical method, simply by changing the phenotype definition.

4.2.1.3 Case-only Modifier Method

Unlike the two previous methods, which use all cases and controls in comparative analyses, the case-only modifier directly compares allele frequencies of two subphenotype groups without use of unaffected individuals. This provides a direct test of phenotypic differences attributable to genetic heterogeneity. This analysis has high power to find genetic risk factors that differ between the two groups. Conversely, it will fail to detect any factors shared

between both groups. Thus, this is strictly a test for heterogeneity of association between genotype and phenotype; it is not a test of overall genetic effect. Ideally, this test will also discover new loci for which there is only an effect in one subgroup (since the other subgroup and controls are theoretically identical at this locus). As this method extracts cases from the whole set of individuals being studied, it can be universally applied regardless of study design.

4.2.1.4 Likelihood Method for Genome-wide Scans

Lee et al. propose a likelihood-based method to test for association, specifically to identify any variants associated with phenotypic heterogeneity [37]. The purpose of this method is to identify multiple types of modifying and subtype-specific variants from a genome-wide scan, rather than improve power for detecting associations via genome-wide scans. This method uses log-linear modeling to test for association in two stages. In the first stage, two models (null and unstructured genetic effect) are compared using a two degree-of-freedom likelihood ratio test. If, and only if, the first test is rejected at the prescribed level, the procedure proceeds to stage two in which multiple models are compared using multiple information criteria to identify exactly what effect the variant (which has already shown some level of association with the unstructured genetic effect model) has on the disease. The models considered in the second stage of the likelihood-based method are *basic*, *subset*, *inv-subset*, *general*, and *modifier* which directly correspond to the five allele frequency possibilities for phenotypic heterogeneity given in Figure 4.1.

After all of these models are fit in stage two, the AIC and BIC of the subtype-specific models are compared. The model with the lowest AIC/BIC classifies the variant. Thus, this approach can detect numerous types of variant associations genome-wide, including the three primary types discussed. The authors do note that while this method is useful in identifying many types of variants, it is not necessarily the most powerful technique for each specific type of variant that may be present, as it employs two-stage testing.

Another potential issue with this likelihood-based approach is that it requires the specification of the population disease subtype frequency, s. If the value of s is misspecified, the

correct variant model may not be chosen. This method also relies on log-linear modeling, so it can only take independent cases and controls. However, extending this modeling approach to accept case-parent trios or other combined data types is feasible using conditional logistic regression or other maximum likelihood approaches.

4.2.1.5 Gene-by-Environment Method

In the case of family-based design (i.e. trios), genetic sources of phenotypic heterogeneity were tested using the genotypic transmission disequilibrium test (gTDT) [70]. This method uses cases and pseudo-controls in a conditional logistic regression framework. Pseudo-controls are created from a trio consisting of two founders and one proband (i.e., affected offspring). Using the genotypes of the founders, all possible genotypes of offspring are calculated. From this list of possible offspring genotypes, the observed genotype of the proband is treated as a case while the unobserved genotypes are used as controls in a conditional logistic regression. Similar to the standard TDT, only informative founder pairs are used for analysis.

To investigate the heterogeneity in association results for two or more phenotypes, the conditional logistic regression models are fit with interactions between genotype and phenotype case status. This provides a measure of association for the effect that case type has in moderating the effect of genotype on affection status.

This method will theoretically identify genetic variants with differing effects between subphenotypes, controlling for parental genotypes. However, this method requires the assumption of a genetic model and has particularly low power to detect and subtype differences, except in very common variants. Thus, it will be outperformed by other methods unless the true underlying genetic model is known.

The technique can be performed on case-parent trios using the R package trio. It is not extendable to any other data structure as the subtype indicator variable is defined at the trio level.

4.2.2 Summary-level Tests

Individual genotype-level data is a gold standard for genetic associations, but frequently only summary statistics from previous association scans are available. There are still useful methods for detecting genetic sources of phenotypic heterogeneity that leverage summary statistics. Furthermore, assessing heterogeneity after individual association scans provides an intuitive and easy to use framework for detecting modifying and subtype-specific variants.

The three methods presented in the following section are capable of detecting subtype-specific and modifier genetic variants as they compare the effect estimates from subtype-specific analyses. It does not detect an overall association, i.e. shared variants. It is important to note that the failure to detect a difference in two subphenotypes does not demonstrate proof that the variant in question is shared between them both.

Summary-level tests are attractive as they can be performed post primary analyses and do not require individual genotypes. Furthermore, they can be used for any statistical test, provided it results in an effect estimate and corresponding standard error, and are not limited to one study design. These methods are generally more flexible than the genotype-level tests.

4.2.2.1 Overlapping Confidence Intervals Method

A naïve, but reasonable, approach to assessing heterogeneity is readily available through summary statistics. Analyzing subgroups separately, one can obtain two effect estimates and their corresponding confidence intervals. These resulting regions are represented by the following:

$$\hat{Q}_1 \pm 1.96 \times S\hat{E}_1$$

$$\hat{Q}_2 \pm 1.96 \times S\hat{E}_2$$

The confidence intervals are then compared. Only if they are disjoint, is there said to be any difference between the effects of the two subphenotypes. The significance level of the confidence intervals can be changed to investigate the evidence of difference between the two subphenotypes. One must be especially careful in the interpretation of visually examining the overlap of confidence intervals. Non-overlapping confidence intervals indicate

statistically significantly different point estimates, whereas overlapping confidence intervals do NOT indicate non-significantly different point estimates.

4.2.2.2 Q Statistic Method

Unlike the visual inspection of confidence intervals, the Q statistic method provides statistical framework for detecting differences in effect estimates [69].

This method is less conservative than the method of examining overlapping confidence intervals, i.e. the overlapping confidence intervals method will reject a null hypothesis of no association every time it is rejected via the Q statistic method, but the converse is not true [69].

The difference in these two methods can be seen by examining the difference intervals.

Q statistic method:

$$\left(\hat{Q}_1 - \hat{Q}_2\right) \pm 1.96\sqrt{\left(\hat{SE_1}^2 + \hat{SE_2}^2\right)}$$

Examining overlap method:

$$\left(\hat{Q}_1 - \hat{Q}_2\right) \pm 1.96\left(\hat{SE_1} + \hat{SE_2}\right)$$

In both methods, the null hypothesis of no difference is rejected when the interval does not contain 0. The difference between the Q statistic method and that of the overlapping confidence intervals is the width of the interval: the interval from the overlapping method is always larger than that of the Q statistic method.

The Q statistic method assumes that estimates are (1) consistent, (2) asymptotically normal, and (3) asymptotically independent [69]. These requirements are typically satisfied in the case of examining log odds ratios from two non-overlapping association scans, given a large enough sample size.

4.2.2.3 Cochran's Q Method

Cochrans Q is a test statistic for assessing heterogeneity of the effects of multiple studies in a meta-analysis setting [14].

The idea of detecting heterogeneity among two or more signals comes from meta-analysis. One main assumption of meta-analysis is that the individual effects are homogeneous. If so, the combined effect estimate is a true representation of the individual signals. Thus interpretation of meta-analysis results depends on identifying heterogeneity, which is commonly tested in a fixed-effects meta-analysis setting using Cochrans Q. This test seeks to find loci for which the individual effects are heterogeneous. This approach is conservative, i.e. it has low power to detect weak heterogeneity, especially when only two individual effects are being tested.

There is also an extension of Cochrans Q for a random-effects meta-analysis; however, in the context of orofacial clefting, we are looking at combining only two or three effects, making a random-effects approach inappropriate. The same philosophical conclusions would also hold for a random-effects test of heterogeneity.

Cochrans Q is calculated in the following way:

The pooled treatment effect is a weighted average of the individual treatment effects (e.g. log odds ratios).

$$T_{pooled} = \frac{\sum w_i T_i}{\sum w_i}$$
$$\text{where } w_i = \frac{1}{SE(T_i)^2)}$$

The standard error of the pooled treatment effect is given by //

$$SE(T_{pooled}) = \frac{1}{\sqrt{\sum w_i}}$$

The Cochran Q statistic is given by

$$Q = \sum_{i=1}^{k} w_i (T_i - T_{pooled})^2$$

which follows a Chi-square distribution with k-1 degrees of freedom.

Cochrans Q statistic, which employs the pooled treatment effect, measures deviation from a weighted average of two estimates. If there is large deviation, we reject the null hypothesis that the effects are the same. This addresses a fundamentally different question

than the one we ask when studying phenotypic heterogeneity. The pooled treatment effect is assumed to be the true underlying genetic effect and only variants with enough distance from this effect show evidence of heterogeneity. Although this formulation of this method does not visually lend itself to the idea of detecting heterogeneity, it can be shown that for the two subgroups, it is identical to the Q-statistic method. Thus, Cochrans Q provides an extension of the Q-statistic method for more than three groups.

This method is applicable in any study design as it is a comparison of summary statistics – it does not require genotype-level data. This method may also be used to compare more than two subgroup estimates; however, it is underpowered to detect a difference in fewer than 5 groups.

Table 4.1: Comparison of the methods for testing genetic sources of phenotypic heterogeneity.

Method	**Type of variants detected**	**Study Design Requirements**
Pooled ($p1 + p2$) v. control	shared	genotype-level data
Separate $p1$ v. control, $p2$ v. control	subtype-specific, modifier*	genotype-level data
Case-only modifier $p1$ v. $p2$	modifier, subtype-specific*	genotype-level data, cases only
Likelihood	subtype-specific, modifier*, shared*	genotype-level data, case-control
GxE	modifier*	genotype-level data, case-parent trios
Overlapping Confidence Intervals	subtype-specific*, modifier*	summary-level data
Q-statistic	subtype-specific*, modifier*	summary-level data
Cochrans Q	subtype-specific*, modifier*	summary-level data

* indicates that this method is not well-powered to find this type of variant

4.3 RESULTS

In order to demonstrate the properties and efficiency of the most commonly used methods for detecting phenotypic heterogeneity, a toy simulation was performed. Genotypes were randomly simulated assuming HWE and some true genetic model: (1) the risk was shared equally among the two case subtypes (i.e. shared), (2) only one case subtype had increased risk (i.e. subtype), (3) the risk increased linearly across the two case subtypes (i.e. gradient), (4) the risk increased for one subtype and decreased for the other (i.e. opposite). The GRRs under which these genotypes were simulated are summarized in Table 4.2.

Table 4.2: Genotypic relative risks for phenotypic heterogeneity demonstration.

Model	Genotypic Relative Risk		
	Controls	Subtype 1	Subtype 2
Shared	1	3	3
Subtype	1	3	1
Modifier - gradient	1	1.5	3
Modifier - opposite	1.5	1	3

For the combined, separate and case-only modifier tests, genotypes of the two case subtypes and controls were directly compared according to the prescribed statistical procedure. Additionally, the two sets of results from the separate analysis for case subtype 1 and case subtype 2 were compared using Cochran's Q and the Q-statistic. The resulting p-values from these association tests are given in Table 4.3.

Table 4.3: Example performance of methods for testing genetic sources of phenotypic heterogeneity under multiple true models.

True Model	Analysis P-value					
	Combined	Separate		Modifier	Cochran's Q	Q-statistic
		type 1 v. control	type 2 v. control			
Shared	3.57×10^{-29}	3.82×10^{-26}	1.15×10^{-22}	3.80×10^{-1}	5.38×10^{-1}	5.38×10^{-1}
Subtype-specific	3.58×10^{-15}	6.42×10^{-30}	3.02×10^{-1}	8.39×10^{-28}	2.41×10^{-11}	2.41×10^{-11}
Modifier (gradient)	1.82×10^{-15}	2.67×10^{-22}	3.28×10^{-5}	6.93×10^{-11}	1.96×10^{-4}	1.96×10^{-4}
Modifier (opposite)	1.66×10^{-2}	3.67×10^{-13}	7.07×10^{-6}	1.25×10^{-26}	7.59×10^{-16}	7.59×10^{-16}

Under the shared genetic model, the combined association clearly outperforms the subtype-specific model. However, with this strong of a genetic effect, it is worthwhile to note that both separate analyses demonstrate rather strong association signals. Appropriately, the case-only modifier and comparison methods do not show evidence of statistically different signals.

The performance of the various association tests under subtype-specific genetic model is also expected; the separate analysis including only those cases with increased risk demonstrated the most statistical significance; but the combined, case-only modifier, and summary-level comparison approaches all detected association.

The gradient genetic model, in which there is increased risk for each subtype but one subtype has increased risk over the other, may be the underlying genetic model that is not completely obvious from the statistical results. In this situation, caution must be exercised in interpreting the results from these multiple association scans. In this situation, examining the GRRs and corresponding confidence intervals may be more illuminating that the association test p-values.

When the effects for the two case subtypes are in the opposite direction, the case-only modifier detects this difference most optimally, followed by the summary-level comparison approaches. Notably in this situation, if the two subtypes were combined for analysis, no association would be detected. This underscores the importance of examining phenotypic heterogeneity not just within the top results from a combined analysis, but also from subtype specific analyses.

4.4 DISCUSSION

There are many possible approaches to examine genetic sources of phenotypic heterogeneity - each with it's own advantages and disadvantages. Examining phenotypic heterogeneity can be done using genotype-level data or summary-level data. As summary statistics are more easily obtainable than genotype-level data, the summary-level tests can provide valuable insight into the genetic architecture of complex traits. These methods tend to be more con-

servative than the genotype-level test, but are useful when genotype level data is unavailable, and when the true genetic model of risk loci is unknown.

Currently, the only method to assess genetic contributions to phenotypic heterogeneity using case-parent trios at the genotype-level is the gene-by-environment framework, which has very low statistical power and is not feasible for small sample sizes nor low-frequency variants. In these situations, and in those with mixed data structures, summary-level methods, which can combined results from multiple analyses including those with different data structures (with the assumption that the effects are the same), are invaluable.

Examining each of these methods together highlights the need for post-statistical testing analysis. Without examining the estimated effects, the underlying genetic model will remain obscured. It is the result of the phenotypic heterogeneity methods coupled with a depiction of the effect estimates for each subtype that will paint a more comprehensive picture of the statistical model of the genetic variants. Doing so may elucidate the biologic mechanism for each risk locus, and in turn, further understanding of the genetic architecture of complex traits.

5.0 IDENTIFYING GENETIC SOURCES OF PHENOTYPIC HETEROGENEITY IN OROFACIAL CLEFTS BY TARGETED SEQUENCING

Jenna C. Carlson[1], Margaret A. Taub[2], Eleanor Feingold[1,3], Terri H. Beaty[4], Jeffrey C. Murray[5], Mary L. Marazita[3,6,7], Elizabeth J. Leslie[7]

[1] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [2] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA. [3] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [4] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore MD, USA. [5] Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, Iowa, 52242, USA. [6] Clinical and Translational Science, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA. [7] Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, 15219, USA.

This chapter has been accepted for publication at Birth Defects Research Part A - Clinical and Molecular Teratology and is currently in press.

5.1 ABSTRACT

Background: Orofacial clefts (OFCs), including nonsyndromic cleft lip with or without cleft palate (NSCL/P), are common birth defects. NSCL/P is highly heterogeneous with multiple phenotypic presentations. Two common subtypes of NSCL/P are cleft lip (CL) and cleft

lip with cleft palate (CLP) which have different population prevalence. Similarly, NSCL/P can be divided into bilateral and unilateral clefts, with unilateral being the most common. Individuals with unilateral NSCL/P are more likely to be affected on the left side of the upper lip, but right side affection also occurs. Moreover, NSCL/P is twice as common in males as in females. The goal of this study is to discover genetic variants that have different effects in case subgroups.

Methods: We conducted both common variant and rare variant analyses in 1,034 individuals of Asian ancestry with NSCL/P, examining four sources of heterogeneity within CL/P: cleft type, sex, laterality, and side.

Results: We identified several regions associated with subtype differentiation – cleft type differences in 8q24 ($p=1.00 \times 10^{-4}$), laterality differences in *IRF6*, a gene previously implicated with wound healing ($p=2.166 \times 10^{-4}$), sex differences and side of unilateral CL differences in *FGFR2* ($p=3.00 \times 10^{-4}$, $p=6.00 \times 10^{-4}$), and sex differences in *VAX1* ($p< 1.00 \times 10^{-4}$) among others.

Conclusions: Many of the regions associated with phenotypic modification were either adjacent to or overlapping functional elements based on ENCODE chromatin marks and published craniofacial enhancers. We have identified multiple common and rare variants as potential phenotypic modifiers of NSCL/P, and suggest plausible elements responsible for phenotypic heterogeneity, further elucidating the complex genetic architecture of OFCs.

5.2 INTRODUCTION

Orofacial clefts (OFCs) are common birth defects, affecting approximately 1 in 800 births worldwide [39]. Approximately 30% of OFCs are syndromic, occurring in combination with some other structural, cognitive, or developmental anomalies. The remaining 70% of OFCs occur as isolated (i.e. nonsyndromic) defects. Nonsyndromic OFCs have complex etiology with multiple genetic and environmental factors interacting to influence risk.

Nonsyndromic OFCs are highly heterogeneous with multiple phenotypic presentations [17]. OFCs are most commonly divided into three major subtypes: cleft lip (CL), cleft

palate (CP), and cleft lip with cleft palate (CLP). CL and CLP share a defect of the lip and are commonly combined for analyses as cleft lip with or without cleft palate (CL/P) [21] [23]. CL/P and CP have historically been considered distinct disorders with separate etiologies because of the different developmental origins of the lip and palate and markedly different prevalence rates in males and females (CP is twice as common in females as in males, while the opposite is true for CL/P [53]). However, they occasionally occur within the same family, an event known as mixed clefting commonly observed in syndromic OFCs, including Van der Woude syndrome [39].

The CL/P subgroup itself is quite heterogeneous and can be further subdivided into bilateral and unilateral clefts, affecting either the left or right side of the upper lip. Of these, left sided unilateral clefts are the most common and bilateral clefts are the least common [26]. The causes of variability in phenotype are largely unknown, and may arise due to underlying genetic factors, different environmental exposures, or other unknown factors. There have been many studies investigating the genetic architecture of NSCL/P, most collapsing cleft subtypes into one larger group (primarily CL/P) for analysis [17] [39]. While this approach is powerful to identify sources of genetic variation that contribute to overall NSCL/P, any signal from genetic variation specific to only one subtype or that differentiates subtypes will be masked. Very few studies have explored genetic associations for clefting phenotypes beyond CL and CLP. There is some evidence that the 13q31 locus near *SPRY2* has a stronger effect in CLP [30] [46]. Similarly, variants in IRF6 are more strongly associated with CL than CLP [52] [63]. Recent evidence suggests that *GREM1* is associated with clefts in the lip and soft palate [47]. Furthermore, variants in *GRHL3* are associated with CP and not with CL/P [42] [50] [78]. Examining CL/P subtypes may elucidate more of the complex genetic architecture of OFCs by identifying genetic mechanisms that modify cleft subtype.

We hypothesized that genetic components of phenotypic heterogeneity, including any contribution of rare variants, can be found for recognized clefting loci. We performed association tests for four sources of phenotypic heterogeneity within CL/P: cleft type (CL vs. CLP), sex (male vs. female), laterality (unilateral vs. bilateral), and side (right unilateral vs. left unilateral) in targeted sequencing from the CleftSeq study [40].

5.3 METHODS

5.3.1 Sample

We compared subtypes within clefting cases from the CleftSeq study to investigate the potential genetic contribution to clefting heterogeneity. CleftSeq is a targeted sequencing study of 13 previously reported loci associated with NSCL/P [40]. These 13 regions, totaling 6.3 Mb, were comprised of 9 high-priority candidates from previous GWAS and/or genome-wide linkage studies and 4 regions containing candidate genes with prior evidence of rare variants contributing to NSCL/P (Table 1.1). Sequencing was performed on 1,498 case-parent trios from Europe, the United States, China and the Philippines.

From the 1,489 trios, we extracted 1,034 probands with NSCL/P of Asian (i.e. Chinese or Filipino) ancestry for analysis and cross-classified them using the four clefting subtype definitions (Table 5.1). Among the 1,034 cases, 33 with unknown laterality were excluded from the analysis of laterality and side of cleft lip groups.

Table 5.1: Sample used for modifier analyses by population.

	Cleft Type		Sex		Laterality		Side of Cleft Lip	
	CL	*CLP*	*Female*	*Male*	*Unilateral*	*Bilateral*	*Right*	*Left*
China	117	284	126	275	278	101	112	166
Philippines	171	462	219	414	440	182	147	293
Total	288	746	345	689	718	283	259	459

5.3.2 Common Variant Analysis

For each factor (i.e. cleft type, sex, laterality, and side), we performed a case vs. case analysis, directly comparing allele frequencies at each SNP between the two groups (e.g. CL vs. CLP, male vs. female, etc.). This type of analysis has very high power to find genetic risk factors that differ between the two groups, but it has no power to find factors that

are important in both groups. Thus this design is strictly a test for heterogeneity in the genotype/phenotype relationship, not an overall test of genetic effect. Ideally, this test will discover new loci for which there is an effect in only one subgroup; such loci may be masked in an overall scan when groups are combined.

We analyzed the association between the four cleft subtype phenotypes and 19,982-20,089 common SNPs (MAF > 0.01) in the thirteen candidate regions by directly comparing the two case subtypes using traditional Chi-Square tests for association. Each Asian population (Chinese and Filipino) was analyzed separately to account for any population stratification. Low-quality SNPs (missing genotypes $> 5\%$ or HWE $p < 0.0001$) were excluded from analyses.

Inverse-variance effects-based meta-analysis of the two population-specific scans was performed on 13,183-13,427 SNPs to detect any signal common to Asian populations. SNPs were excluded from the meta-analyses if they were flagged as low-quality in at least one population-specific analysis, or if effects were heterogeneous between populations (Cochrans Q $p < 0.05$). Statistical significance was determined using a Bonferroni threshold adjusting for four scans of thirteen regions of 9.615×10^{-4} (i.e. $0.05/52$). This threshold allows for the generation of hypotheses regarding the genetic mechanisms of clefting subtypes and thus is not as strictly conservative as a Bonferroni correction for the number of markers tested (5200 tests, p-value threshold of 1×10^{-5} [40]). Thus, the suggestive associations found in this study should be followed up rigorously. Common variant analyses were performed using PLINK software [61].

5.3.3 Rare Variant Analysis

Rare variants (MAF < 0.01) were also interrogated for association with subtype differentiation using the same phenotype definitions as in the common variant analysis.

First, variants within exons of canonical transcripts of each gene were examined using gene-based versions of the Collapsed Multivariate and Combining (CMC) test [44] and the Sequence Kernel Association Test (SKAT) [82].

Secondly, two window-based approaches were used to investigate burdens of all rare variants. SNPs were combined into regions using two window-based methods – 2,662 windows using a fixed window size of 5Kb with 2.5Kb overlap between windows, and 14,232 windows using exactly 20 SNPs per window with 10 SNP overlap between windows (windows at the end of each region contained at least 14 SNPs). Each window was comprised of SNPs from only one of the candidate regions. Windows are highly correlated within each candidate region, so statistical significance was again determined using a Bonferroni threshold of 9.615×10^{-4}. Rare variants were analyzed with the SKAT option in RVTESTS software [83].

5.3.4 Functional Annotation of Rare Variant Windows

The CleftSeq project sequenced 6.3Mb of largely non-coding DNA around these GWAS and OFC candidate genes. We failed to identify significant associations in analyses of coding variants (results not shown), so we hypothesized that functional variants would be regulatory. We examined intervals containing overlapping windows for functional elements based on ENCODE chromatin marks [15] [67] and published craniofacial enhancers [3] [10] [62].

5.4 RESULTS

5.4.1 Cleft Type

In the common variant meta-analysis, 20 SNPs from 4 loci were significantly associated with CL v. CLP differentiation (Figure 5.1A). These associations were seen in SNPs on 9q22 near *PTCH1* and *FOXE1*, on 17p22 near *NOG*, and on 20q12 near *MAFB*. Specifically, a set of variants in and near *PTCH1* were more strongly associated with CL than with CLP (lead SNP: rs202111971 $p = 6.484 \times 10^{-4}$, Figure 5.1B). A neighboring set of variants did not show formally significant differences by cleft type, but tended to be more strongly associated with CLP (Figure 5.1B). In the 9q22 region, a set of SNPs downstream of the *FOXE1* transcription start site were more strongly associated with CLP than with CL (lead SNP: rs73492791 $p = 1.138 \times 10^{-4}$, Figure 5.1C). Moreover, minor alleles in the 17p22 regions and 20q12 regions

were more strongly associated with CLP (lead SNPs: rs7208145 $p = 9.041 \times 10^{-4}$, rs6129626 $p = 5.039 \times 10^{-4}$, Figure 5.1C,E). Notably, none of these SNPs associated with cleft type differentiation (CL vs. CLP) was significantly associated with risk of OFC overall [40].

Twenty-five windows of rare variants in the *PAX7*, *ARHGAP29*, 8q24, *FOXE1*, *VAX1*, *NTN1*, and *NOG* sequencing regions were significantly associated with cleft type differentiation (CL vs. CLP) (Table D1). Of these, two sets of three overlapping windows (8:129790677-129795772 [min $p = 4.50 \times 10^{-4}$] and 8:130298273-130305772 [min $p = 1.00 \times 10^{-4}$]) on 8q24 are particularly interesting because they contain SNPs that individually show strong association with NSCL/P in Europeans. Furthermore, one of these intervals (8:129790677-129795772) consisting of three overlapping windows was located adjacent to a putative regulatory element as defined by H3K27Ac marks in multiple cell types from ENCODE (Figure 5.4A).

5.4.2 Laterality

In the common variant meta-analysis, 27 SNPs from 2 loci were significantly associated with laterality differences (Figure 5.2A). These associations were seen for 26 SNPs on 1q32 near *IRF6* (lead SNP: rs6540559 $p = 2.166 \times 10^{-4}$, Figure 5.2B) and a single SNP on 17p22 near *NOG* (rs184942776 $p = 5.262 \times 10^{-5}$, Figure 5.2C). SNPs in *IRF6* were associated with differentiation between bilateral and unilateral CL/P. Specifically, minor alleles of SNPs in *IRF6* were associated with unilateral CL/P. The minor alleles at these SNPs also are significantly protective against overall OFC risk (Table D5).

Differences in CL/P laterality were observed in 17 windows of rare variants (Table 5.1). Despite having many overlapping windows of rare variants, there was no evidence of known regulatory or enhancer elements within these intervals.

5.4.3 Sex

While no significant associations for sex differences were observed in the common variant analysis (Figure 5.3A), 28 windows of rare variants were significantly associated with sex differences (Table D3).

Eight windows defining three larger intervals (10:118624030-118629029 [min p = 5.00×10^{-4}], 10:118638519-118644029 [min p < 1.00×10^{-4}], and 10:118851530-11885725 [min p < 1.00×10^{-4}]) near *VAX1* were significantly associated with sex differences in Filipinos. One of these intervals (10:118851530-11885725), comprised of three windows near *VAX1*, overlapped a craniofacial regulatory element identified from p300 ChIP-Seq in craniofacial tissue in mouse embryos [3] [77] (Figure 5.4B). It is unclear what gene is regulated by this element, as the activity pattern of the enhancer resembles the endogenous expression of both adjacent genes *VAX1* and *SHTN1* [2] [16]. Interestingly, other significant windows in this region occurred immediately downstream of *SHTN1*.

Two non-overlapping windows near *FGFR2* (10:123368869-123373868 [p = 3.00×10^{-4}] and 10:123479803-123483275 [p = 8.00×10^{-4}]) were also significantly associated with sex differences in Filipinos. The first of these windows overlapped multiple regulatory annotations including a binding site for p63, a transcription factor known to regulate *FGFR2* [20] [22] (Figure 5.4C). The second window overlaps more regulatory annotations characteristic of epithelial enhancers (Figure 5.4C).

5.4.4 Side of Lip

We did not observe any significant associations with right unilateral vs. left unilateral CL/P in the common variant analysis (Figure 5.3B). However, 13 windows of rare variants were significantly associated with side of cleft lip differentiation (Table D4). Interestingly, one window near *FGFR2* (10:123431369-123436368 [p = 6.00×10^{-4}]) was significantly associated with side of cleft lip differentiation in Filipinos and was adjacent to active enhancers from human neural crest cell lines and a putative palate enhancer from p300 ChIP-seq of mouse palatal tissue (Figure 5.4C).

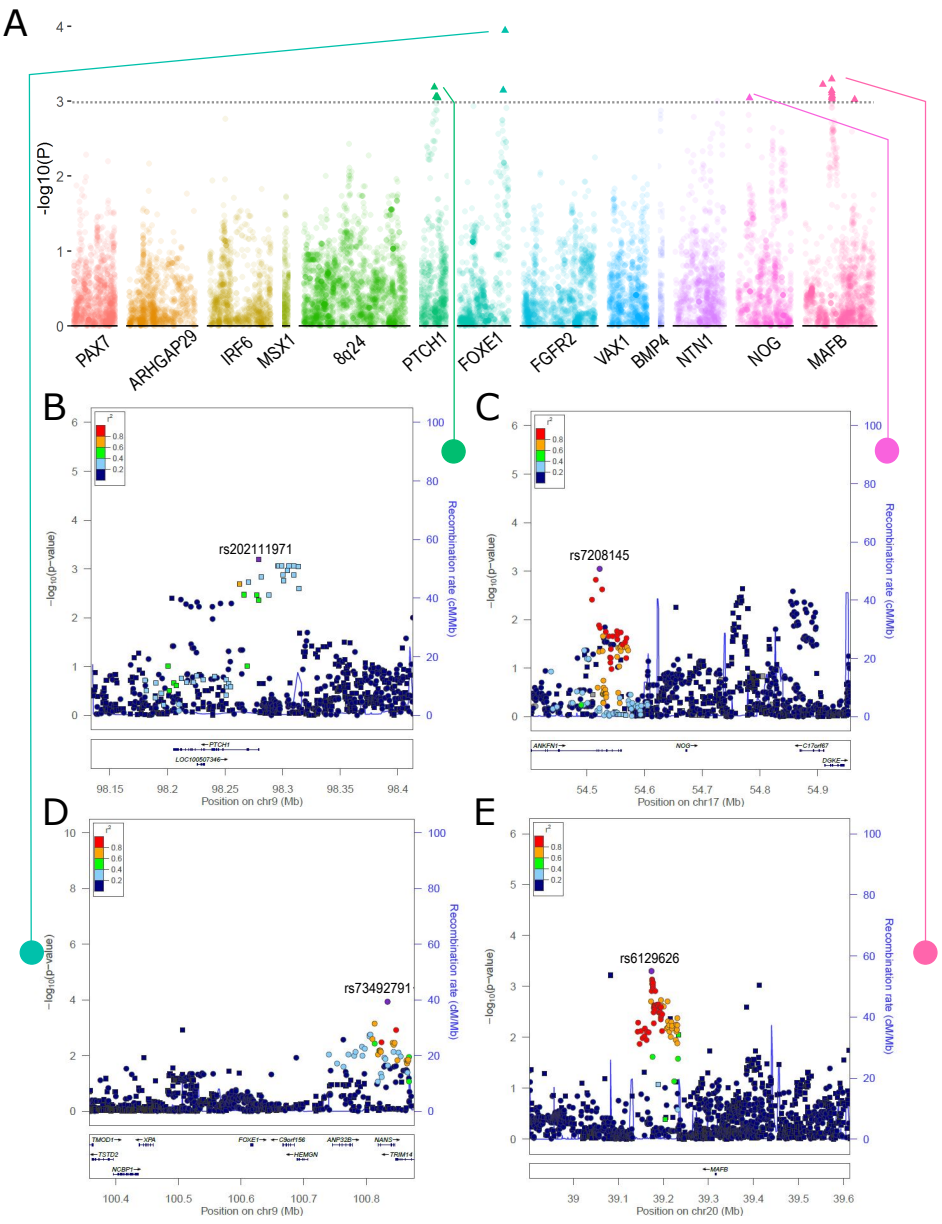


Figure 5.1: CL vs. CLP cleft type modifiers. (A) Cleft type (CL vs. CLP) association results from the common-variant meta-analysis of Filipino and Chinese populations. (B)-(E) Regional association plots for 9q22 (x2), 17q22, and 20q12 showing log10(P-values) for SNPs with stronger association with CL (squares) and stronger association with CLP (circles) based on the direction of the odds ratio. Plots were generated using LocusZoom [60]. The recombination overlay (blue line, right y-axis) indicates the boundaries of the LD-block. Points are color coded according to pairwise linkage disequilibrium (r2) with the index SNP.

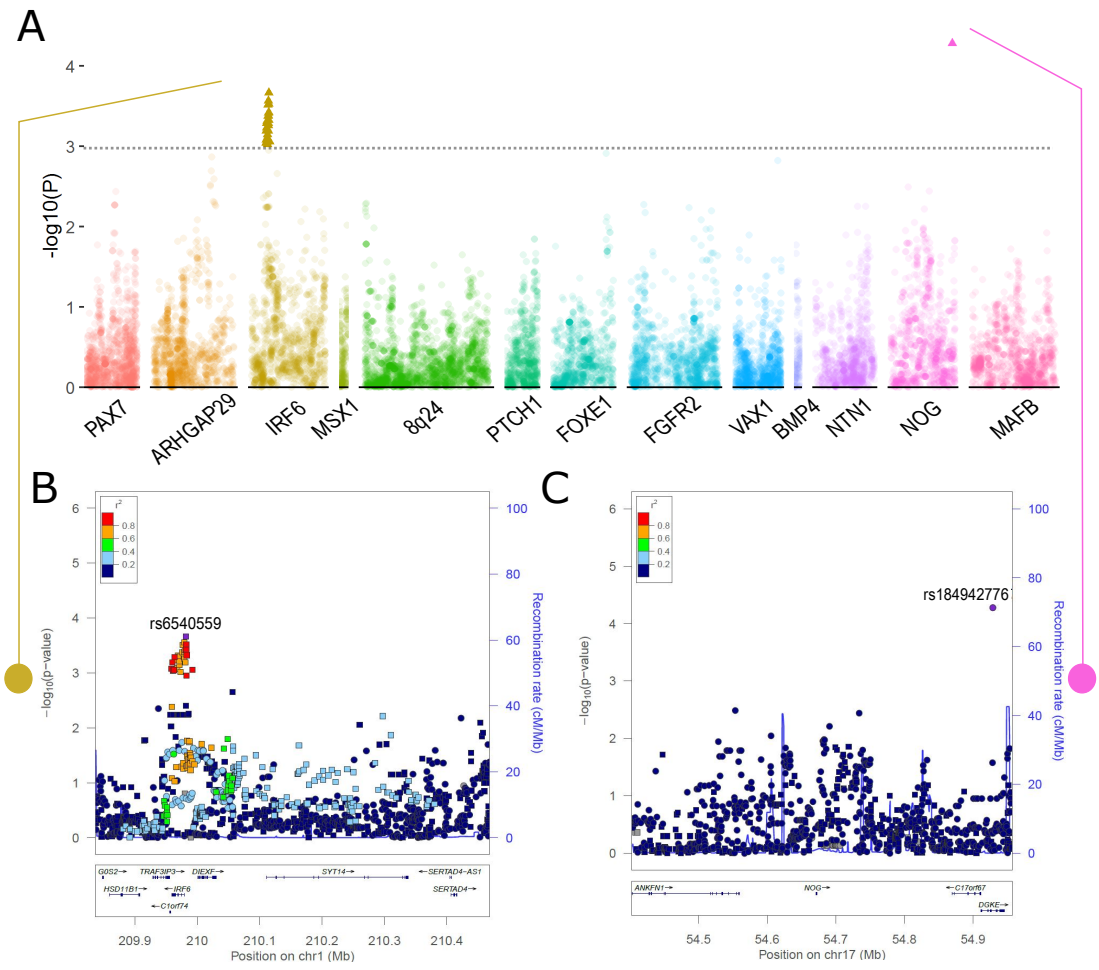


Figure 5.2: Unilateral vs. bilateral CL/P modifiers. (A) Laterality (unilateral vs. bilateral) association results from the common-variant meta-analysis of Filipino and Chinese populations. (B) Regional association plots for 1q32 showing log10(P-values) for SNPs with stronger association with unilateral CL/P (squares) and stronger association with bilateral CL/P (circles) based on the direction of the odds ratio. (C) Regional association plots for 17q22 showing log10(P-values) for SNPs with stronger association with unilateral CL/P (squares) and stronger association with bilateral CL/P (circles) based on the direction of the odds ratio. Plots were generated using LocusZoom [60]. The recombination overlay (blue line, right y-axis) indicates the boundaries of the LD-block. Points are color coded according to pairwise linkage disequilibrium (r2) with the index SNP.

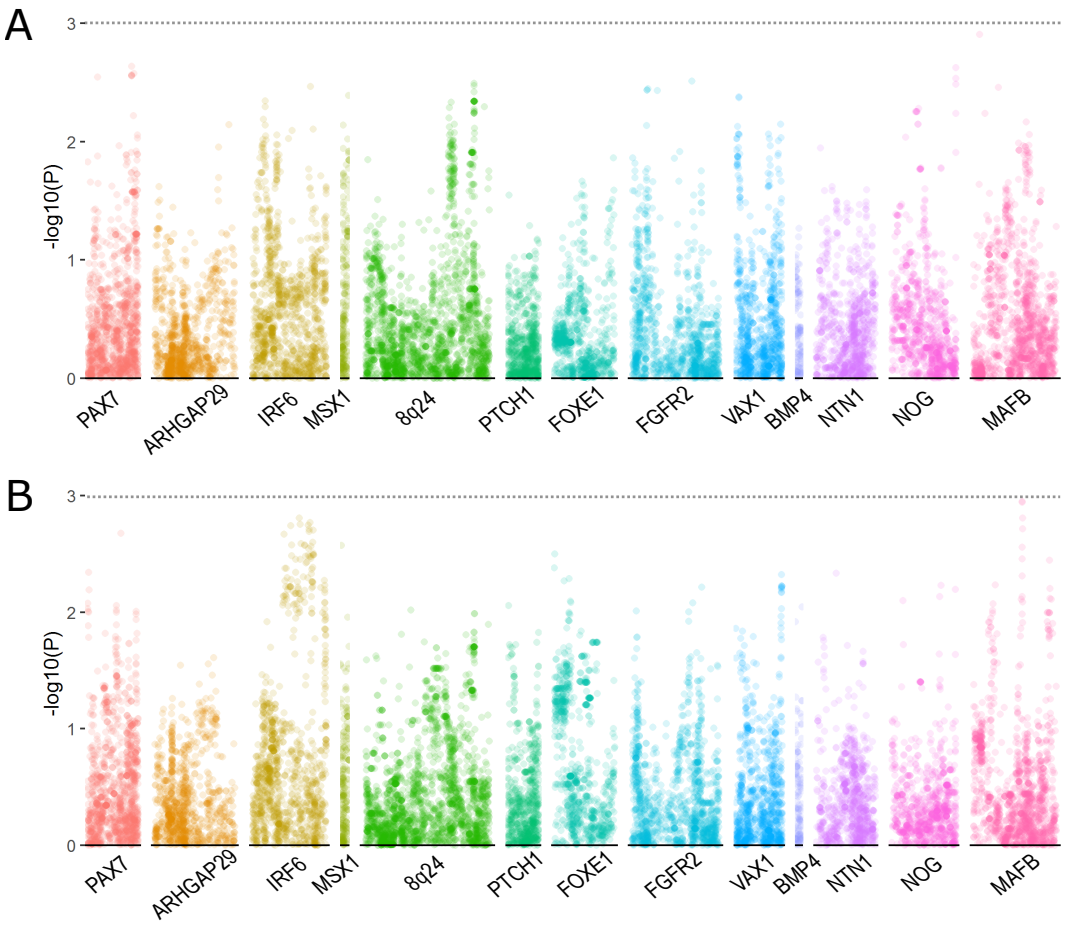


Figure 5.3: Sex-specific modifiers of CL/P. (A) Sex (male vs. female) association results from the common-variant meta-analysis of Filipino and Chinese populations. (B) Side (right unilateral vs. left unilateral) association results from the common-variant meta-analysis of Filipino and Chinese populations.

Figure 5.4: Significant rare variant windows with potential regulatory effects. (A) 8q24 for cleft type, (B) *VAX1* for sex, and (C) *FGFR2* for sex and side.

5.5 DISCUSSION

NSCL/P is a complex disorder with many different anatomical forms. GWASs have identified dozens of genetic associations with NSCL/P [4] [41] [46] [49]; however, a small number of studies have identified cleft subtype specific associations, most of which are reflect differences between CL and CLP [46] [52] [63]. The current study adds to these findings by identifying both common and rare variants that are associated with subtype differentiation in cleft type, laterality, sex, and side of unilateral CL. We performed common and rare variant association testing with four cleft subtypes (cleft type: CL vs. CLP; laterality: unilateral vs. bilateral; sex; and side: right vs. left CL/P) to further interrogate OFC-associated regions from the CleftSeq targeting sequencing study. We identified several regions associated with cleft subtype differentiation – common variants in *IRF6* and rare variants in 8q24, *FGFR2*, and *VAX1*, among others. Notably, these associations are found with both previously known clefting-associated variants and variants that were not significantly associated with overall clefting (CL/P). Multiple associations with regulatory (non-coding) elements and differences in clefting subtypes were discovered, contributing to the evidence that non-coding variants have a significant role in the genetic causes of NSCL/P [39] [40] [63]. However, it is not clear from the association results which alleles are relevant to these phenotypes; systematic studies in model systems will likely be required to identify functional SNPs and a possible mechanism.

We identified 26 SNPs within *IRF6* associated with differences between unilateral and bilateral CL/P. Specifically, individuals with unilateral CL/P had higher frequencies of minor alleles in these 26 variants than did bilateral CL/P individuals. *IRF6* has been previously implicated in wound healing [8] [7] [32], so these cleft laterality differences are particularly interesting. The same alleles showing a protective effect for overall cleft risk were more strongly associated with unilateral CL/P than bilateral. If we consider unilateral CL/P as a less severe presentation of clefting than bilateral CL/P, our finding that OFC-protective variants are associated more strongly with unilateral CL/P and previous evidence that IRF6 is associated with CL [63] together suggest that the *IRF6* locus is associated with decreased risk of severe clefting.

Rare variants on 8q24 were found to significantly differ between CL and CLP, including an interval adjacent to a putative regulatory element. This provides strong evidence for a regulatory role of variants within 8q24 on the presentation of NSCL/P. Furthermore, SNPs on 8q24 have previously shown very strong association with cleft risk in European GWAS [5] [9] [24] [55], but are not associated with cleft risk in Asian GWAS(i.e. in common variant analyses). This may be due to population-specific differences in SNP informativeness within 8q24, which reflects haplotype diversity [55]. SNPs within 8q24 have markedly higher heterozygosity in Europeans than Asians, making common-variant associations within this region far more powerful among Europeans. We hypothesize that this region also is associated with clefting risk in other populations, although the statistical evidence from analyses of common variants is lacking. The association with cleft type differentiation within windows of 8q24 rare variants observed in the Filipino population here may be evidence that some SNPs within 8q24 confer clefting risk in Asian populations.

Additionally, rare variant associations with potential regulatory elements were observed when examining sex differences and markers near *VAX1* and *FGFR2* and those near *FGFR2* and the left vs. right side of unilateral CL/P. While it is not immediately clear how *VAX1* and *FGFR2* specifically contribute to sex differences in NSCL/P, biological hypotheses regarding sex differences in other disorders (e.g. autism) involve a multiple-threshold multifactorial liability model in which females have a higher threshold than males. In other words, affected females are hypothesized to carry a higher mutational burden than affected males. The same would hold for NSCL/P, where there are more affected males than females. Under this hypothesis, relatives of affected females are at increased risk for CL/P, which is supported by population-based recurrence risk estimates from Denmark [25]. A similar threshold model may also pertain to differences in laterality and severity of NSCL/P.

Contrary to the common disease-common variant hypothesis, we observed clear contributions from both common and rare variants in this study of the genetic underpinnings of NSCL/P and the potential differences within NSCL/P subtypes. This work adds to a growing body of evidence implicating rare variants in risk of NSCL/P [1] [38] [40]. Importantly, this work highlights the impact of rare variants as potential phenotypic modifiers, an area that needs larger studies in additional populations that are expanded to the entire

genome. As costs of whole genome sequencing decrease, these studies will be more feasible for NSCL/P and will continue to improve our understanding of the genetic architecture of NSCL/P.

6.0 DETECTING SUBTYPE-SPECIFIC EFFECTS IN OROFACIAL CLEFTING THROUGH GENOME-WIDE ASSOCIATION

Jenna C. Carlson[1], John R. Shaffer[2], Lina Moreno[3], George L. Wehby[4], Eleanor Feingold[1,2], Seth M. Weinberg[5], Jeffrey C. Murray[6], Terri H. Beaty[7], Mary L. Marazita[2,5,8], Elizabeth J. Leslie[5].

[1] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [2] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, USA. [3] Department of Orthodontics, College of Dentistry, University of Iowa, Iowa City, IA 52242, USA. [4] Department of Health Management and Policy, College of Public Health, University of Iowa, Iowa City, IA 52242 USA. [5] Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, PA, 15219, USA. [6] Department of Pediatrics, Carver College of Medicine, University of Iowa, Iowa City, Iowa, 52242, USA. [7] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205 USA. [8] Clinical and Translational Science, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA.

This chapter is part of a manuscript currently being prepared for publication.

6.1 INTRODUCTION

Orofacial clefting is a common complex birth defect with multiple phenotypic presentations. OFCs can arise when there is a disruption in fetal craniofacial development, approximately

during weeks four through ten of embryogenesis [17]. The upper lip and primary palate are formed by the sixth week and the secondary palate is formed by the tenth week [17]; any disruption in these processes many results in an orofacial cleft. Numerous genetic studies (including genome-wide linkage and association) have made substantial progress in identifying genetic risk factors for OFCs in the past decade.

The primary focus of the OFC genetics literature has been on the two most common presentations: cleft lip with or without cleft palate (CL/P) and isolated cleft palate (CP) [17] [39]. CL/P and CP have historically been considered distinct disorders due to the different developmental origins of the lip and palate [31], different prevalence rates among males and females [53], and different proportions of syndromic cases (50% CP vs. 30% for CL/P) [39]. At least 20 genetic risk loci have been identified for CL/P [43]; only one locus has been identified for CP [42]. Despite this progress, the identified risk loci only account for a modest portion of the genetic variance of OFCs, suggesting that additional genetic risk factors may be involved.

In the current study, we sought to identify additional genetic risk variants for specific OFC subtypes - CL, CLP and CP - including exploring the possibility of shared etiology between two or more subtypes. To do so, we conducted genome-wide meta-analyses for CL, CLP, and CP using two large OFC studies.

6.2 METHODS

6.2.1 Contributing GWAS studies

Two consortia contributed to this study (Table 6.1). The first, hereafter called GENEVA OFC, used a family-based design and included 461 case-parent trios with CL, 1143 case-parent trios with CLP, and 451 case-parent trios with CP, from populations in Europe (Denmark and Norway), the United States, and Asia (Singapore, Taiwan, Philippines, Korea, and China). The specifics of this study were previously described in [4] [5]. Briefly, samples were genotyped for 589,945 SNPs on the Illumina Human610-Quadv.1_B Bead-

Chip, genetic data were phased using SHAPEIT, and imputation was performed with IMPUTE2 software to the 1000 Genomes Phase 1 release (June 2011) reference panel. Genotype probabilities were converted to most-likely genotype calls with the GTOOL software (http://www.well.ox.ac.uk/ cfreeman/software/gwas/gtool.html), using a genotype probability threshold of 0.9.

The second consortium included samples contributing to the Pittsburgh Orofacial Cleft (POFC) study, comprising 179 cases and 271 case-parent trios with CL, 644 cases and 1048 case-parent trios with CLP, 78 cases and 165 case-parent trios with CP, plus 1700 unaffected controls. Participants were recruited from 13 countries in North America (United States), Central or South America (Guatemala, Argentina, Colombia, Puerto Rico), Asia (China, Philippines), Europe (Denmark, Turkey, Spain), and Africa (Ethiopia, Nigeria). Additional details on recruitment, genotyping, and quality controls are described in [41] [42]. Briefly, samples were genotyped for 539,473 SNPs on the Illumina HumanCore + Exome array. Data were phased with SHAPEIT2 and imputed using IMPUTE2 to the 1000 Genomes Phase 3 release (September 2014) reference panel and converted to most-likely genotypes for statistical analysis.

A total of 412 individuals were in both the GENEVA OFC and POFC studies, so we excluded these participants from the GENEVA OFC study for this analysis. Informed consent was obtained for all participants and all sites had both local IRB approval and approval at the University of Pittsburgh, the University of Iowa, or Johns Hopkins University.

Table 6.1: Counts of Cases, Controls, and Trios from the POFC and GENEVA studies.

Study		CL		CLP		CP	
	Controls	Trios	Cases	Trios	Cases	Trios	Cases
POFC	1700	271	179	1048	644	165	78
GENEVA	–	461	–	1143	–	451	–

6.2.2 SNP selection

Quality control procedures were completed in each contributing study and have been described extensively in the original publications [41] [42] [4] [5]. In the POFC study, SNPs with minor allele frequencies (MAF) less than 1% or those deviating from HWE (p<0.0001) in genetically defined, unrelated European controls were excluded.

Similarly, SNPs with MAF <1% or those deviating from HWE were excluded. To account for different marker sets and identifiers between the two imputed datasets, the final analysis included only those overlapping SNPs that were matched on chromosome, nucleotide position, and alleles. A total of 6,090,031 SNPs were included in the meta-analysis.

6.2.3 Statistical Analysis

We identified three analysis groups from the contributing studies: a case-control subgroup from POFC, an unrelated case-parent trio group from POFC, and an unrelated case-parent trio group from GENEVA OFC. In the casecontrol subgroup, logistic regression was used to test for association under the additive genetic model, while including 18 principal components of ancestry (generated via principal component analysis [PCA] of 67,000 SNPs in low linkage disequilibrium across all ancestry groups) to adjust for population structure [41]. The two case-parent trio subgroups from POFC and GENEVA were analyzed separately using the TDT. The resulting effect estimates for the three analysis groups were combined in an inverse variance-weighted fixed-effects meta-analysis. The combined estimate, a weighted log odds ratio, was compared to a Chi-squared distribution with two degrees of freedom. This procedure was followed for three GWASs, one for each cleft type.

From these three scans, SNPs demonstrating suggestive association (i.e. $p < 1.00 \times 10^{-5}$) in any of the three scans were considered for further analysis. For each SNP, the effects of CL were compared to those of CLP, and the effects of CLP compared to those of CP using the Q-statistic [69]. These two contrasts were chosen based on the biologic plausibility of shared genetic effects between clefts affecting the lip (CL and CLP) and clefts affecting the palate (CLP and CP). A strict statistical significance threshold was set at 8.6×10^{-4} [(i.e. $0.05/(29 * 2)$], but results were also considered for suggestive evidence. The goal of this

method was to examine the results holistically to gain further understanding of cleft-specific signals. Further, the direction of association was determined by the difference in absolute values of the log odds ratios (i.e. $|log(OR_{CLP})| - |log(OR_{CL})|$, $|log(OR_{CLP})| - |log(OR_{CP})|$).

This set of 1,375 SNPs was collapsed into 29 loci based on genomic position and linkage disequilibrium. Loci spanning large regions with evidence of multiple, statistically independent signals (i.e. *IRF6* and 8q24) were subdivided into multiple groups of SNPs based on LD grouping in PLINK software [41] [61]. The p-values comparing CL to CLP and CLP to CP were averaged across the SNPs within a locus. Similarly, the direction of association for CL to CLP and CLP to CP was averaged across the SNPs within a locus. The sign of this average direction indicated the cleft type with the strongest association signal for that locus.

These loci were then represented graphically on what is hereinafter referred to as the cleft map. The x-axis on the cleft map is given by the average log10 p-value of the comparison between CL and CLP of the locus times the sign of the average direction (CLP or CL) of the locus. The y-axis of the cleft map is given by the average log10 p-value of the comparison between CLP and CP of the locus times the sign of the average direction (CLP or CP) of the locus. Thus, loci nearest the origin do not demonstrate any subtype-specific signals in our sample. Loci along the x-axis in the left half of the map demonstrate evidence for CL-specific association; loci along the y-axis in the lower half of the map demonstrate evidence for CP-specific association; loci in the upper-right quadrant demonstrate evidence for CLP-specific association; loci along the y-axis in the upper half of the map demonstrate evidence for CL/P-specific association. A summary of this is represented in Figure 6.1. Loci further away from the origin exhibit more statistical evidence of cleft-specific signals. Concentric circles about the origin based on log10 p-values of the Q-statistics are given for reference. Also, the size of the point on the map represents the strength of association in the separate GWASs.

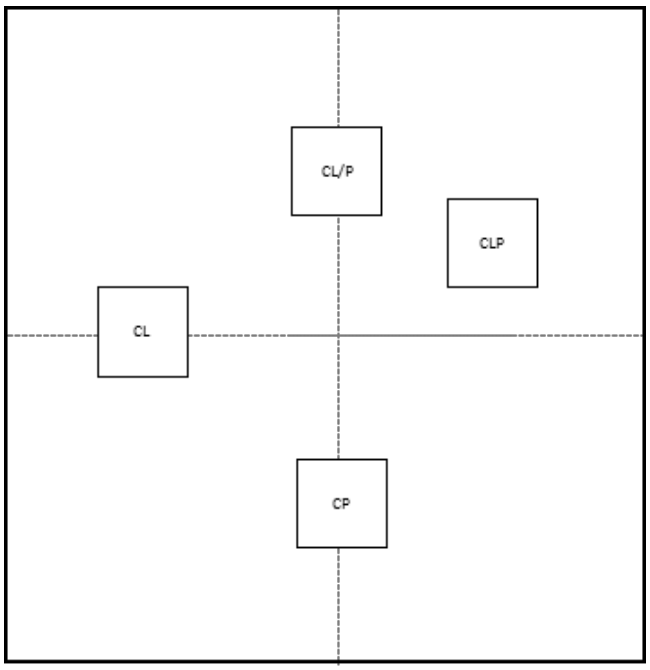


Figure 6.1: Guide to interpret cleft subtype-specific signals

6.3 RESULTS

In our comparison of the effects for CL, CLP, and CP for 29 loci with marginal association for at least one subtype, we identified many loci with stronger association in one cleft subtype in addition to a handful of loci demonstrating no evidence of subtype-specific signals. Specifically, two genes (*UGT3A2* and *GRHL3*) show evidence of CP-specific association (average p-values: ($p_{CLP.CL} = 0.93$, $p_{CLP.CP} = 8.9 \times 10^{-5}$) and ($p_{CLP.CL} = 0.64$, $p_{CLP.CP} = 2.2 \times 10^{-5}$) respectively). We also identified two possible CL-specific associations in *SLC28A3* and possibly *COL8A1* (average p-values: ($p_{CLP.CL} = 4.1 \times 10^{-5}$, $p_{CLP.CP} = 0.24$) and ($p_{CLP.CL} = 4.9 \times 10^{-3}$, $p_{CLP.CP} = 0.011$) respectively). *WNT5A* and *MSX2* are among a few genes demonstrating CLP-specific association (average p-values: ($p_{CLP.CL} = 9.8 \times 10^{-5}$, $p_{CLP.CP} = 0.073$) and ($p_{CLP.CL} = 3.8 \times 10^{-3}$, $p_{CLP.CP} = 3.7 \times 10^{-4}$) respectively). Further, many loci including known CL/P risk genes with substantial evidence

(*IRF6a,b* and 8q24a,b,c gene desert) appear in the combined CL and CLP area (average p-values shown in Table 6.2).

A brief summary of the findings is given in this Table 6.2.

Table 6.2: Average p-values from the Q-statistic comparison of CLP to CL, and CLP to CP for each locus.

Locus	# SNPs	$P_{CLP.CL}$	$P_{CLP.CP}$	Locus	# SNPs	$P_{CLP.CL}$	$P_{CLP.CP}$
PAX7	60	0.084	0.056	*FOXE1*	36	0.67	0.25
CAPZB	21	0.065	2.1×10^{-3}	*VAX1*	41	0.58	0.26
GRHL3	17	0.64	2.2×10^{-5}	*KRT18*	7	0.18	4.5×10^{-5}
ARHGAP29	26	0.79	0.031	*SPRY2*	15	0.16	2.3×10^{-4}
WNT5A	14	9.8×10^{-5}	0.073	*ARID3B*	148	0.82	7.6×10^{-3}
ERC2	18	0.064	0.23	*NTN1*	48	0.24	3.0×10^{-3}
COL8A1	226	4.9×10^{-3}	0.011	*GOSR2*	62	0.50	0.38
TP63	9	0.78	0.013	*NOG*	4	0.47	0.41
SHROOM3	36	0.95	6.3×10^{-4}	*MAFB*	46	0.22	0.011
UGT3A2	7	0.93	8.9×10^{-5}	*IRF6a*	126	0.81	6.2×10^{-8}
MSX2	20	3.8×10^{-3}	3.7×10^{-4}	*IRF6b*	60	0.063	3.0×10^{-4}
TRIM10	1	0.33	0.36	8q24a	136	0.57	3.0×10^{-5}
DCAF4L2	8	0.55	0.023	8q24b	117	0.78	3.1×10^{-3}
BAALC	5	0.012	0.14	8q24c	52	0.022	0.042
SLC28A3	9	4.1×10^{-5}	0.24				

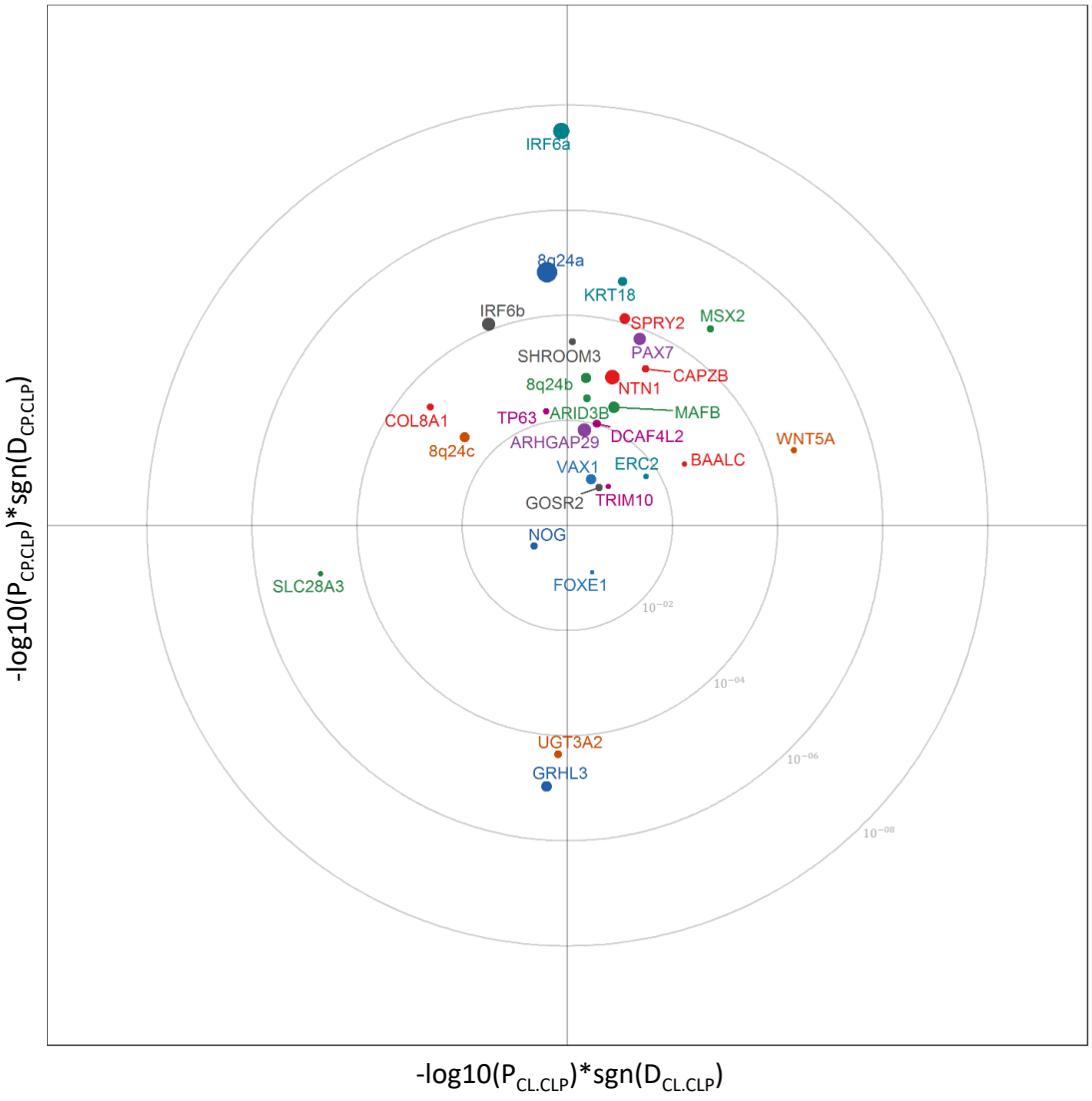


Figure 6.2: Cleft Map

6.4 DISCUSSION

This analysis comparing association signals from three GWAS of the primary cleft subtypes (CL, CLP, and CP) detected numerous cleft-type-specific signals. These findings add to the evidence that many OFC-risk genes operate in a way that may increase risk of one cleft

type. Specifically, we have demonstrated that many known cleft-risk regions (e.g. *IRF6* and the 8q24 gene desert) are common to CL/P as demonstrated by numerous GWASs. Further, we note that *GRHL3*, which is the only gene implicated for isolated cleft palate in a genome-wide association study, demonstrated a CP-specific signal in our analysis. We also note that many candidate genes for OFCs (e.g. *FOXE1*) may generally increase risk of clefting in a non-cleft-type-specific manner. Replication attempts for these and the rest of the regions depicted in the "map" are currently underway to compare our findings to those using a similar procedure in a set of independent OFC cases and controls.

It is important to note the limitations of using statistical comparisons to identify these cleft-type-specific signals. A failure to demonstrate a statistically significant signal for one cleft type does not prove that the signal is common to all clefts. Further investigation of the signals that appear to be common to OFCs in general, regardless of cleft type, should be conducted.

Methods similar to the one proposed here are necessary to explain the system by which OFCs occur. Identifying statistically different association signals among clefting subtypes will be crucial in understanding the biologic systems through which OFCs develop. These methods are fundamental to understanding difference in association signals from multiple subtypes. The naive approach to compare results from separate analyses lacks statistical rigor and is inappropriate. Statistical significance (i.e. p-values) are dependent upon sample size and, as with most clefting studies, there is not perfect balance among the cases of different cleft types; the naive approach (i.e. comparing p-values between association scans) is unsuitable to detect differences.

This method of comparison may also be extended to study other subphenotypes in orofacial clefting or other complex traits. Specifically within clefting, the contribution of subclinical phenotypes to the transmission of cleft-risk variants has been suggested, especially in the case of the *obicularis oris* muscle defects [57]. Extending the subphenotype definitions for OFCs to include the presence and absence of subclinical phenotypes may further demonstrate the role of subclinical phenotypes in the OFC-transmission. Furthermore, this method may be useful in untangling the mechanisms determining cleft severity, which can be measured by the completeness of a cleft and whether one, or both, sides of the face

is affected. The POFC study has detailed phenotyping for individuals with OFC, which when accompanied with this method to distinguish specific signals, may further elucidate the genetic mechanisms operating to form each specific cleft type.

7.0 CONCLUSIONS

In this multifaceted dissertation, we examined two hallmarks of common, complex traits – the analysis of mixed data structures and phenotypic heterogeneity – in a genome-wide association setting. For the analysis of mixed data structures, we provided an overview of existing methods for combining data from case-parent trios and unrelated cases and controls and proposed a new empirical Bayes-type shrinkage estimator for estimating genotypic relative risk in mixed data structures. The application of these methods was also studied in an example of a multiethnic study of orofacial clefts. Furthermore, we examined sources of phenotypic heterogeneity in complex traits, and conducted a philosophical evaluation of existing statistical methods capable of identifying genetic sources of phenotypic heterogeneity. We then applied some of these tools to identify cleft subtype-specific signals in a targeted sequencing study of orofacial clefting and to classify associated variants by which cleft types they are associated with in the context of genome-wide meta-analysis. Through these investigations we have provided insight into the statistical methods commonly employed to address mixed data structures and phenotypic heterogeneity as well as the genetic architecture of orofacial clefting.

7.1 STRENGTHS, LIMITATIONS, AND FUTURE WORK

7.1.1 Empirical Bayes-Type Estimator

The empirical Bayes-type estimator defined in chapter 3 demonstrated no gain in efficiency, likely due to the use of a robust variance estimate, as the analytical variance of the estimator

cannot be formulated. In general, empirical Bayes-type estimators are good ways to gain efficiency even when assumptions are violated, but in this case the maximum likelihood estimates were not tractable. Thus, this method requires a robust sandwich estimate of variance. Notably, conservative estimates were also obtained for the constrained and unconstrained estimators employing robust variance estimates.

As evidenced in the application to the multiethnic study of orofacial clefts, the empirical Bayes-type estimator preserves the ranking of the most significantly-associated regions. This demonstrates the utility of the empirical Bayes-type estimator, in that it appropriately orders genetic variants by strength of association. Further simulation is warranted to investigate if the preservation of ranked variants holds in general.

Additionally, the variance of the empirical Bayes-type estimator could be estimated empirically through permutations. Using an permutation-based variance estimate may improve the behavior of the association test by eliminating the overly conservative nature of the test statistic. By using an empirical variance estimate, the resulting test statistic should more closely follow a uniform distribution under the null hypothesis and improve the behavior of the significance test. Thus, this empirical Bayes-type estimator proposed in chapter 3 may demonstrate practical importance to mixed data studies, as it preserves the order of significant variants and could be improved through empirically estimated variance.

This method assumes an additive genetic model for each variant. This does not present a problem in studying OFCs, as the majority of the variants discovered do not deviate severely from an additive model. However, should other genetic models be desired, this method could be extended to model them. This method currently assesses association for common variants only; low-frequency variants are excluded because of the nature of the estimator. There are existing rare variant methods that work for case-control and trios separately and SKAT for larger families, but none that incorporates mixed data types [28] [73]. Development of a similar, empirical Bayes-types estimator to assess a burden of rare variants would require defining the constrained and unconstrained models; the notation of HWE across a set of SNPs would need to be specified.

Many extensions of the TDT, which account for missing parental genotypes, an additional offspring, and affected parents exist. Currently, this method does not provide similar

accommodations, but extensions may be feasible. Along those same lines, the only family-based data this method takes is the case-parent trio. It may be beneficial to extend this method to larger pedigrees by including them in the formation of the likelihood. This would require specifying the exact nature of the likelihood, and would not necessarily account for all familial relationships like some genome-wide association methods do.

This method does not adjust for principal components of ancestry or other covariates, but rather assumes homogeneity of effects within the population being studied.

Lastly, the construction of the partial M-estimator assumes $\hat{\theta}$ and V_B are known quantities. It could be adapted to incorporate the estimation of these parameters into the estimating equations at the expense of computational time.

7.1.2 Phenotypic Heterogeneity

We explored many new aspects for detecting genetic differences for trait differences in complex disease. We presented numerous statistical methods for detecting genetic variants associated with phenotypic heterogeneity, both for analyzing genotype-level data and for combining summary-level results from multiple analyses. In doing so, we identified the situations in which each method for detecting phenotypic heterogeneity is most optimal.

Additionally, in chapter 5, we determined specific genetic loci associated with differentiating OFC subphenotypes in a targeted sequencing study. We observed both common and rare variants contributing to the genetic underpinnings and subphenotype differences of CL/P. Importantly, this work motivates the study of rare variants as potential phenotypic modifiers.

Furthermore, we developed a novel visualization tool for displaying genetic risk factors for OFCs in chapter 6. This tool presents statistical evidence for heterogeneity within a risk locus by examining differences in effects between multiple subphenotypes of OFCs. Particularly, this tool avoids the traditional dichotomization of statistical significance (i.e. p-value < 0.05) but rather visualizes the statistical evidence of heterogeneity for each genetic locus. From this visualization, we can motivate potential biologic mechanisms which may drive the heterogeneous process through which OFCs arise.

Future analyses examining phenotypic heterogeneity will include examining gene×gene interactions between loci shown as potential modifiers with CL/P risk loci or general OFC risk loci. Ideally, these potential interactions will motivate biological mechanisms underlying the genetic association with OFCs.

The analysis of genetic sources of phenotypic heterogeneity may provide insight into the mechanisms by which OFCs come about. Examining genetic sources of phenotypic heterogeneity may be excellent for nominating possible biologic mechanisms/processes that may be responsible for modifying phenotypes. Moreover, the methods and tools discussed here can and should be applied to the study of other complex traits to advance understanding of the genetic architecture of diverse traits across the phenotypic spectrum.

APPENDIX A

TABLE OF ABBREVIATIONS

Table A1: Commonly-used abbreviations

GWAS	Genome-wide association study
SNP	single-nucleotide polymorphism
OFC	orofacial cleft
CL	isolated cleft lip
CLP	cleft lip and palate
CL/P	cleft lip with/without cleft palate
CP	isolated cleft palate
NSCL/P	nonsyndromic cleft lip with/without cleft palate
TDT	transmission-disequilibrium test
CPG	conditional on parental genotype
GRR	genotypic relative risk
MAF	minor allele frequency
HWE	Hardy-Weinberg Equilibrium

APPENDIX B

VARIANCE CALCULATION FOR THE EMPIRICAL BAYES-TYPE ESTIMATOR

B.1 ROBUST SANDWICH ESTIMATE OF VARIANCE FOR CONSTRAINED ESTIMATE

$$V_N(\hat{\beta}^0) = \frac{1}{N}\sum_{1}^{N}\frac{(f_{1i}(\beta^0)x_i + f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))^2\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)^2}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)^2\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)^2} \tag{B.1.1}$$

$$f_{1i} = \left(\frac{N_{1i}}{p_{1i}}\right)\left(\frac{\delta p_{1i}(\beta)}{\delta\beta}\right)$$

$$f_{2,\boldsymbol{gpi},i} = \left(\frac{N_{\boldsymbol{gpi},i}}{p_{\boldsymbol{gpi},i}}\right)\left(\frac{\delta p_{\boldsymbol{gpi},i}}{\delta\beta}\right)$$

B.2 ROBUST SANDWICH ESTIMATE OF VARIANCE FOR UNCONSTRAINED ESTIMATE

$$V_N(\hat{\beta}^0) =$$
$$\frac{1}{N}\sum_{1}^{N}\frac{(f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))^2\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)^2}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)^2\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)^2} \tag{B.2.1}$$

$$f_{1i} = \left(\frac{N_{1i}}{p_{1i}}\right)\left(\frac{\delta p_{1i}(\beta)}{\delta\beta}\right)$$
$$f_{2,\boldsymbol{gpi},i} = \left(\frac{N_{\boldsymbol{gpi},i}}{p_{\boldsymbol{gpi},i}}\right)\left(\frac{\delta p_{\boldsymbol{gpi},i}}{\delta\beta}\right)$$

B.3 ROBUST SANDWICH ESTIMATE OF VARIANCE FOR EMPIRICAL BAYES-TYPE ESTIMATE

$$V_N(\hat{\beta}_{EB}) =$$

$$\frac{1}{N}\sum_{1}^{N}(f_{1i}(\beta^0)x_i + f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))(-\boldsymbol{K}(\beta-\beta^0)+\beta-\beta_{EB})\times$$

$$\frac{\left(\boldsymbol{K}x_i f'_{1i}(\beta) - \boldsymbol{K}x_i f'_{2,\boldsymbol{gpi},i}(\beta) + \boldsymbol{K}f'_{2,\boldsymbol{gpi},i}(\beta)\right)}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)}$$

$$+\left(\boldsymbol{K}x_i f'_{1i}(\beta) - \boldsymbol{K}x_i f'_{2,\boldsymbol{gpi},i}(\beta) + \boldsymbol{K}f'_{2,\boldsymbol{gpi},i}(\beta)\right)\times$$

$$\frac{(f_{1i}(\beta^0)x_i + f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))(-\boldsymbol{K}(\beta-\beta^0)+\beta-\beta_{EB})}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)}+ \tag{B.3.1}$$

$$\frac{f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))^2\left(\boldsymbol{K}x_i f'_{1i}(\beta) - \boldsymbol{K}(1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)^2\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)}- \tag{B.3.2}$$

$$\frac{(\boldsymbol{K}-1)(f_{1i}(\beta^0)x_i + f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))(f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)^2}$$

$$-\frac{(\boldsymbol{K}-1)((f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))(-\boldsymbol{K}(\beta-\beta^0)+\beta-\beta_{EB}))}{x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)}$$

$$+\frac{(\boldsymbol{K}-1)(f_{1i}(\beta^0)x_i + f_{2,\boldsymbol{gpi},i}(\beta^0)(1-x_i))(f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))}{\left(x_i f'_{1i}(\beta^0) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta^0)\right)\left(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)\right)^2}\times$$

$$(\boldsymbol{K}x_i f'_{1i}(\beta) + \boldsymbol{K}(1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta))$$

$$-\frac{(\boldsymbol{K}-1)^2(f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))^2}{(x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta))^2}$$

$$-\frac{(\boldsymbol{K}-1)(f_{1i}(\beta)x_i + f_{2,\boldsymbol{gpi},i}(\beta)(1-x_i))(-\boldsymbol{K}(\beta-\beta^0)+\beta-\beta_{EB})}{x_i f'_{1i}(\beta) + (1-x_i)f'_{2,\boldsymbol{gpi},i}(\beta)}$$

$$+(-\boldsymbol{K}(\beta-\beta^0)+\beta-\beta_{EB})^2 \tag{B.3.3}$$

$$\text{where } f_{1i} = \left(\frac{N_{1i}}{p_{1i}}\right)\left(\frac{\delta p_{1i}(\beta)}{\delta\beta}\right)$$

$$f_{2,\boldsymbol{gpi},i} = \left(\frac{N_{\boldsymbol{gpi},i}}{p_{\boldsymbol{gpi},i}}\right)\left(\frac{\delta p_{\boldsymbol{gpi},i}}{\delta\beta}\right) \text{ and}$$

$$\boldsymbol{K} = \hat{V}_{\hat{\beta}}(\hat{V}_{\hat{\beta}} + \hat{\theta}^2\hat{\Delta}^T\hat{\Delta})^{-1}$$

APPENDIX C

SUPPLEMENTAL FIGURES

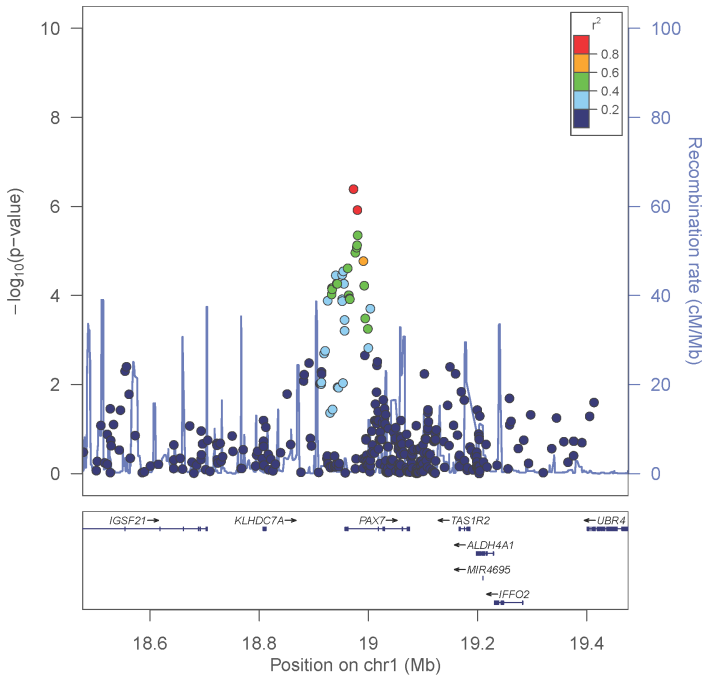


Figure C1: Regional association plot showing $-log10(P-value)$ for genotyped SNPs at the 1p36 locus from the meta-analysis of the European-ancestry group

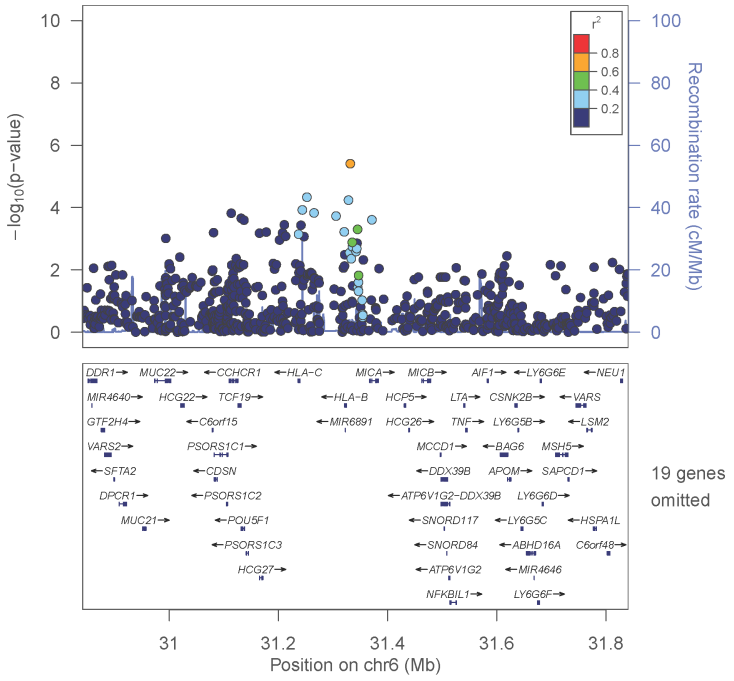


Figure C2: Regional association plot showing $-log10(P-value)$ for genotyped SNPs at the 6p21 locus from the meta-analysis of the European-ancestry group

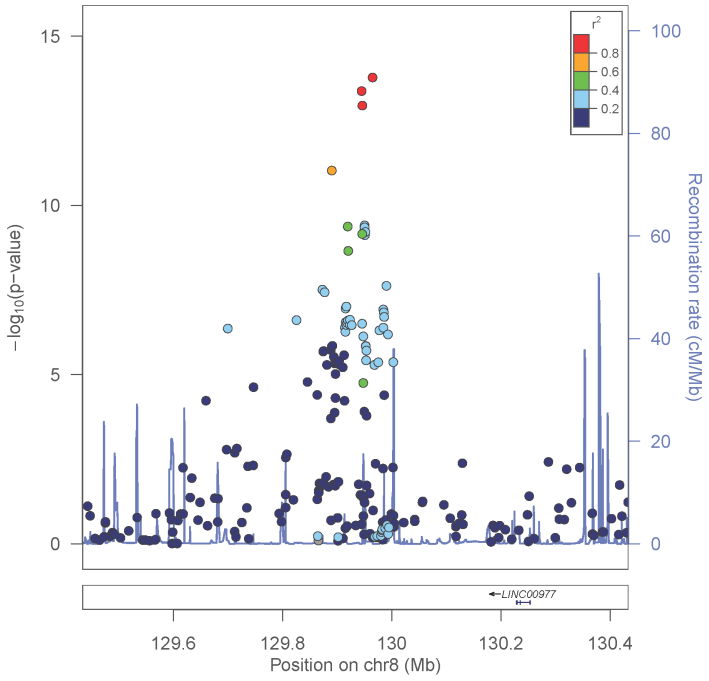


Figure C3: Regional association plot showing $-log10(P-value)$ for genotyped SNPs at the 8q24 locus from the meta-analysis of the European-ancestry group

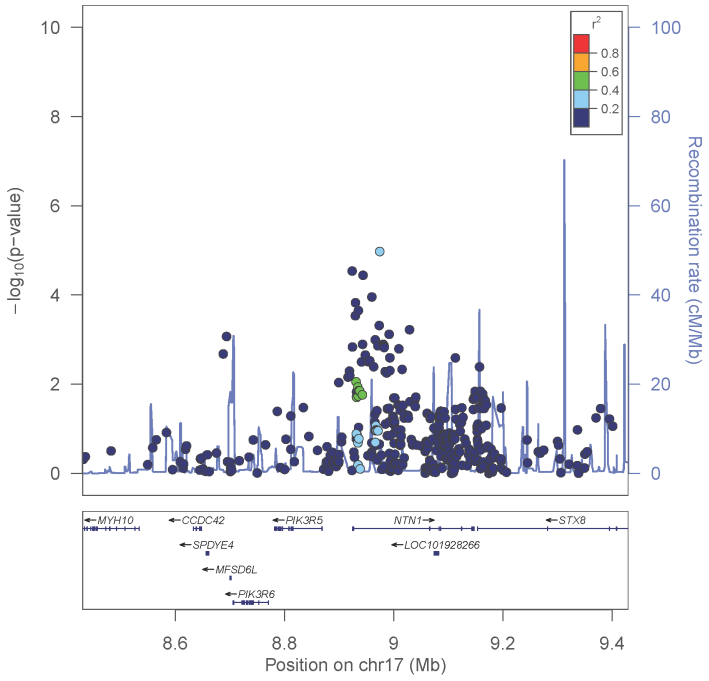


Figure C4: Regional association plot showing $-log10(P-value)$ for genotyped SNPs at the 17p13 locus from the meta-analysis of the European-ancestry group

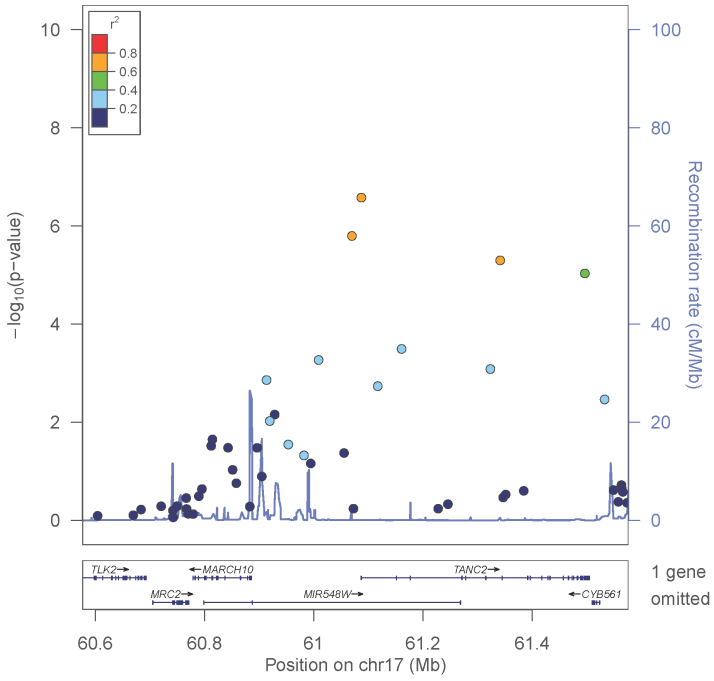


Figure C5: Regional association plot showing $-log10(P-value)$ for genotyped SNPs at the 17q23 locus from the meta-analysis of the European-ancestry group

APPENDIX D

SUPPLEMENTAL TABLES

Table D1: Windows of rare variants with statistically significant association with cleft type (CL vs. CLP).

Window Range	Population	Window Type	P-value
1:18943576-18946340	Philippines	20 SNP	6.00E-04
1:18999691-19004690	Philippines	5 KB	6.00E-04
1:19000487-19002797	Philippines	20 SNP	4.00E-04
1:19002191-19007190	Philippines	5 KB	3.00E-04
1:19002383-19004086	Philippines	20 SNP	3.00E-04
1:19004691-19009690	Philippines	5 KB	9.00E-04
1:19005587-19007798	Philippines	20 SNP	5.00E-04
1:95007579-95012578	China	5 KB	2.00E-04
8:129790677-129793236	Philippines	20 SNP	8.00E-04
8:129790773-129795772	Philippines	5 KB	6.00E-04
8:129791975-129794554	Philippines	20 SNP	4.50E-04
8:130298273-130303272	Philippines	5 KB	1.00E-04
8:130300154-130302665	Philippines	20 SNP	2.50E-04
8:130300773-130305772	Philippines	5 KB	5.00E-04
9:100464682-100465908	China	20 SNP	2.00E-04
9:100487529-100492528	China	5 KB	3.00E-04
9:100680029-100685028	Philippines	5 KB	2.00E-04
9:100681080-100683520	Philippines	20 SNP	8.00E-04
10:118459030-118464029	Philippines	5 KB	7.00E-04
10:118653761-118655989	China	20 SNP	3.00E-04
10:118724030-118729029	Philippines	5 KB	6.00E-04
17:8998160-9001204	China	20 SNP	7.50E-04
17:9000016-9005015	China	5 KB	1.00E-04
17:9057516-9062515	China	5 KB	7.00E-04
17:54817004-54818659	China	20 SNP	8.00E-04

Table D2: Windows of rare variants with statistically significant association with laterality (unilateral vs. bilateral).

Window Range	Population	Window Type	P-value
1:94925079-94930078	Philippines	5 KB	6.00E-04
1:94925714-94928273	Philippines	20 SNP	4.00E-04
1:94927110-94929746	Philippines	20 SNP	9.00E-04
1:94981785-94983950	Philippines	20 SNP	4.50E-04
8:129515829-129517924	Philippines	20 SNP	7.00E-04
8:129557483-129559460	China	20 SNP	6.00E-04
8:129575655-129577099	China	20 SNP	1.00E-04
8:129576195-129578952	China	20 SNP	4.00E-04
8:129888186-129889673	China	20 SNP	< 0.0001
9:98133816-98138815	Philippines	5 KB	7.00E-04
9:98136401-98138871	Philippines	20 SNP	7.00E-04
9:98148816-98153815	China	5 KB	7.00E-04
9:98149174-98150718	China	20 SNP	1.00E-04
17:9234571-9236098	China	20 SNP	6.50E-04
17:54567773-54572772	China	5 KB	1.00E-04
20:39260700-39265699	China	5 KB	3.00E-04
20:39264089-39265758	China	20 SNP	1.00E-04

Table D3: Windows of rare variants with statistically significant association with sex (male vs. female).

Window Range	Population	Window Type	P-value
1:19198100-19199678	Philippines	20 SNP	2.00E-04
1:19199721-19200817	Philippines	20 SNP	7.00E-04
1:94327579-94332578	China	5 KB	5.00E-04
1:94328844-94331028	China	20 SNP	5.00E-05
1:94342579-94347578	China	5 KB	9.00E-04
1:94385079-94390078	China	5 KB	8.00E-04
4:4876429-4878242	China	20 SNP	9.50E-04
8:130028642-130030604	China	20 SNP	3.50E-04
9:100697529-100702528	Philippines	5 KB	9.00E-04
9:100697713-100700826	Philippines	20 SNP	9.00E-04
10:118541530-118546529	Philippines	5 KB	3.00E-04
10:118596325-118598622	Philippines	20 SNP	4.00E-04
10:118624030-118629029	Philippines	5 KB	5.00E-04
10:118625260-118628030	Philippines	20 SNP	7.00E-04
10:118638519-118640461	Philippines	20 SNP	1.00E-04
10:118639030-118644029	Philippines	5 KB	2.00E-04
10:118639485-118641706	Philippines	20 SNP	< 0.0001
10:118851530-118856529	Philippines	5 KB	2.00E-04
10:118854083-118856514	Philippines	20 SNP	< 0.0001
10:118854900-118857525	Philippines	20 SNP	1.00E-04
10:123368869-123373868	Philippines	5 KB	3.00E-04
10:123479803-123483275	Philippines	20 SNP	8.00E-04
17:54615714-54618317	China	20 SNP	6.00E-04
17:54766942-54767309	Philippines	20 SNP	9.50E-04
17:54811309-54812853	Philippines	20 SNP	4.00E-04

Table D3 (continued)

Window Range	Population	Window Type	P-value
20:39291288-39293331	China	20 SNP	9.00E-04
20:39444857-39446029	China	20 SNP	1.00E-04
20:39445740-39446718	China	20 SNP	6.00E-04

Table D4: Windows of rare variants with statistically significant association with side (unilateral left vs. unilateral right).

Window Range	Population	Window Type	P-value
1:19044691-19049690	Philippines	5 KB	3.00E-04
1:19044786-19046526	Philippines	20 SNP	2.50E-04
1:94745898-94747471	China	20 SNP	8.00E-04
1:94860079-94865078	Philippines	5 KB	7.00E-04
1:94861794-94864557	Philippines	20 SNP	5.00E-04
9:98306107-98308002	Philippines	20 SNP	6.50E-04
9:100631558-100633036	Philippines	20 SNP	5.00E-04
9:100652529-100657528	China	5 KB	9.00E-04
10:123431369-123436368	Philippines	5 KB	6.00E-04
17:54567379-54569266	Philippines	20 SNP	2.00E-04
17:54567773-54572772	Philippines	5 KB	6.00E-04
17:54568409-54570881	Philippines	20 SNP	4.00E-04
20:39231214-39232819	Philippines	20 SNP	7.00E-04

Table D5: Modifier association results for laterality (unilateral vs. bilateral) and TDT results for NSCL/P for variants within *IRF6*.

CHR	SNP	BP	Minor Allele	Major Allele	OR (modifier)	P-value (modifier)	OR (TDT)	P (TDT)
1	rs4844895	209958580	C	T	0.6869	8.32E-04	0.6541	2.34E-10
1	rs2235372	209960436	A	G	0.6753	6.46E-04	0.6558	9.56E-10
1	rs742214	209960925	C	T	0.6899	9.13E-04	0.6697	1.78E-09
1	rs742215	209961023	A	T	0.6896	9.06E-04	0.6721	2.56E-09
1	rs2073485	209962794	A	G	0.6891	8.89E-04	0.6697	1.97E-09
1	rs2235373	209963803	A	G	0.6766	5.18E-04	0.6613	5.96E-10
1	rs2235375	209965587	C	G	0.691	5.91E-04	0.6912	1.44E-08
1	rs6685182	209968319	A	C	0.698	8.31E-04	0.6973	3.35E-08
1	rs2013162	209968684	A	C	0.6869	4.84E-04	0.6954	2.48E-08
1	rs2236907	209971628	A	C	0.6947	7.18E-04	0.6991	4.10E-08
1	rs2236908	209971640	C	G	0.6872	5.07E-04	0.6907	1.39E-08
1	rs2236909	209971655	G	A	0.6917	6.15E-04	0.6972	3.03E-08
1	rs2294408	209973549	A	G	0.7005	9.52E-04	0.6943	2.32E-08
1	rs2073486	209976215	A	G	0.6837	4.22E-04	0.6872	9.20E-09
1	rs2073487	209976646	C	T	0.6776	3.07E-04	0.6942	2.09E-08
1	rs17015250	209978777	G	T	0.6887	5.19E-04	0.6888	1.02E-08

Table D5 (continued)

CHR	SNP	BP	Minor Allele	Major Allele	OR (modifier)	P-value (modifier)	OR (TDT)	P (TDT)
1	rs12403599	209979014	C	G	0.6895	5.52E-04	0.6831	5.83E-09
1	rs7545538	209979613	G	C	0.6832	4.16E-04	0.6912	3.05E-08
1	rs7545542	209979635	T	C	0.6744	2.75E-04	0.6774	3.63E-09
1	rs2357229	209980489	T	G	0.6801	3.85E-04	0.6798	5.18E-09
1	rs1005287	209980757	A	G	0.6928	6.48E-04	0.6849	7.20E-09
1	rs6540559	209982025	A	G	0.6572	2.17E-04	0.6606	7.01E-10
1	rs6696825	209982372	G	A	0.6715	3.80E-04	0.6679	1.42E-09
1	rs6659367	209982408	T	G	0.6649	3.00E-04	0.6667	1.97E-09
1	rs764093	209983331	G	A	0.6765	4.77E-04	0.6631	6.77E-10
1	rs12070337	209992127	A	G	0.6901	8.82E-04	0.6679	1.42E-09

BIBLIOGRAPHY

[1] T. Al Chawa, K. U. Ludwig, H. Fier, B. Potzsch, R. H. Reich, G. Schmidt, B. Braumann, N. Daratsianos, A. C. Bohmer, H. Schuencke, M. Alblas, N. Fricker, P. Hoffmann, M. Knapp, C. Lange, M. M. Nothen, and E. Mangold. Nonsyndromic cleft lip with or without cleft palate: Increased burden of rare variants within gremlin-1, a component of the bone morphogenetic protein 4 pathway. *Birth Defects Res A Clin Mol Teratol*, 2014.

[2] C. Armit, S. Venkataraman, L. Richardson, P. Stevenson, J. Moss, L. Graham, A. Ross, Y. Yang, N. Burton, J. Rao, B. Hill, D. Rannie, M. Wicks, D. Davidson, and R. Baldock. emouseatlas, emage, and the spatial dimension of the transcriptome. *Mamm Genome*, 23(9-10):514–24, 2012.

[3] C. Attanasio, A. S. Nord, Y. Zhu, M. J. Blow, Z. Li, D. K. Liberton, H. Morrison, I. Plajzer-Frick, A. Holt, R. Hosseini, S. Phouanenavong, J. A. Akiyama, M. Shoukry, V. Afzal, E. M. Rubin, D. R. FitzPatrick, B. Ren, B. Hallgrimsson, L. A. Pennacchio, and A. Visel. Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, 342(6157):1241006, 2013.

[4] T. H. Beaty, J. C. Murray, M. L. Marazita, R. G. Munger, I. Ruczinski, J. B. Hetmanski, K. Y. Liang, T. Wu, T. Murray, M. D. Fallin, R. A. Redett, G. Raymond, H. Schwender, S. C. Jin, M. E. Cooper, M. Dunnwald, M. A. Mansilla, E. Leslie, S. Bullard, A. C. Lidral, L. M. Moreno, R. Menezes, A. R. Vieira, A. Petrin, A. J. Wilcox, R. T. Lie, E. W. Jabs, Y. H. Wu-Chou, P. K. Chen, H. Wang, X. Ye, S. Huang, V. Yeow, S. S. Chong, S. H. Jee, B. Shi, K. Christensen, M. Melbye, K. F. Doheny, E. W. Pugh, H. Ling, E. E. Castilla, A. E. Czeizel, L. Ma, L. L. Field, L. Brody, F. Pangilinan, J. L. Mills, A. M. Molloy, P. N. Kirke, J. M. Scott, M. Arcos-Burgos, and A. F. Scott. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near mafb and abca4. *Nature Genetics*, 42(6):525–9, 2010.

[5] T. H. Beaty, I. Ruczinski, J. C. Murray, M. L. Marazita, R. G. Munger, J. B. Hetmanski, T. Murray, R. J. Redett, M. D. Fallin, K. Y. Liang, T. Wu, P. J. Patel, S. C. Jin, T. X. Zhang, H. Schwender, Y. H. Wu-Chou, P. K. Chen, S. S. Chong, F. Cheah, V. Yeow, X. Ye, H. Wang, S. Huang, E. W. Jabs, B. Shi, A. J. Wilcox, R. T. Lie, S. H. Jee, K. Christensen, K. F. Doheny, E. W. Pugh, H. Ling, and A. F. Scott. Evidence for

gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genetic Epidemiology*, 35(6):469–78, 2011.

[6] N. W. Berk and M. L. Marazita. Costs of cleft lip and palate: Personal and societal implications. In D. F. Wyszynski, editor, *Cleft Lip and Palate: From Origin to Treatment*, pages 458–467. Oxford University Press, Inc., New York, New York, 2002.

[7] L. C. Biggs, L. Rhea, B. C. Schutte, and M. Dunnwald. Interferon regulatory factor 6 is necessary, but not sufficient, for keratinocyte differentiation. *Journal of Investigative Dermatology*, 132(1):50–8, 2012.

[8] L. C. Biggs, R. L. Naridze, K. A. DeMali, D. F. Lusche, S. Kuhl, D. R. Soll, B. C. Schutte, and M. Dunnwald. Interferon regulatory factor 6 regulates keratinocyte migration. *J Cell Sci*, 127(Pt 13):2840–8, 2014.

[9] S. Birnbaum, K. U. Ludwig, H. Reutter, S. Herms, M. Steffens, M. Rubini, C. Baluardo, M. Ferrian, N. Almeida de Assis, M. A. Alblas, S. Barth, J. Freudenberg, C. Lauster, G. Schmidt, M. Scheer, B. Braumann, S. J. Berge, R. H. Reich, F. Schiefke, A. Hemprich, S. Potzsch, R. P. Steegers-Theunissen, B. Potzsch, S. Moebus, B. Horsthemke, F. J. Kramer, T. F. Wienker, P. A. Mossey, P. Propping, S. Cichon, P. Hoffmann, M. Knapp, M. M. Nothen, and E. Mangold. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genetics*, 41(4):473–7, 2009.

[10] J. F. Brinkley, S. Fisher, M. P. Harris, G. Holmes, J. E. Hooper, E. W. Jabs, K. L. Jones, C. Kesselman, O. D. Klein, R. L. Maas, M. L. Marazita, L. Selleri, R. A. Spritz, H. van Bakel, A. Visel, T. J. Williams, J. Wysocka, C. FaceBase, and Y. Chai. The facebase consortium: a comprehensive resource for craniofacial researchers. *Development*, 143 (14):2677–88, 2016.

[11] L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–9, 2001.

[12] H. Chen, C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T. Sofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedon, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet*, 98(4): 653–66, 2016.

[13] Y. H. Chen and H. W. Lin. Simple association analysis combining data from trios/sibships and unrelated controls. *Genet Epidemiol*, 32(6):520–7, 2008.

[14] W. G. Cochran. The comparison of percentages in matched samples. *Biometrika*, 37 (3-4):256–66, 1950.

[15] E. P. Consortium, R. Myers, J. Stamatoyannopoulos, M. Snyder, I. Dunham, R. Hardison, B. E. Bernstein, T. Gingeras, W. Kent, and E. Birney. A user's guide to the encyclopedia of dna elements (encode). *PLoS Biology*, 9(4):e1001046, 2011.

[16] G. Diez-Roux, S. Banfi, M. Sultan, L. Geffers, S. Anand, D. Rozado, A. Magen, E. Canidio, M. Pagani, I. Peluso, N. Lin-Marq, M. Koch, M. Bilio, I. Cantiello, R. Verde, C. De Masi, S. A. Bianchi, J. Cicchini, E. Perroud, S. Mehmeti, E. Dagand, S. Schrinner, A. Nurnberger, K. Schmidt, K. Metz, C. Zwingmann, N. Brieske, C. Springer, A. M. Hernandez, S. Herzog, F. Grabbe, C. Sieverding, B. Fischer, K. Schrader, M. Brockmeyer, S. Dettmer, C. Helbig, V. Alunni, M. A. Battaini, C. Mura, C. N. Henrichsen, R. Garcia-Lopez, D. Echevarria, E. Puelles, E. Garcia-Calero, S. Kruse, M. Uhr, C. Kauck, G. Feng, N. Milyaev, C. K. Ong, L. Kumar, M. Lam, C. A. Semple, A. Gyenesei, S. Mundlos, U. Radelof, H. Lehrach, P. Sarmientos, A. Reymond, D. R. Davidson, P. Dolle, S. E. Antonarakis, M. L. Yaspo, S. Martinez, R. A. Baldock, G. Eichele, and A. Ballabio. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biology*, 9(1):e1000582, 2011.

[17] M. J. Dixon, M. L. Marazita, T. H. Beaty, and J. C. Murray. Cleft lip and palate: understanding genetic and environmental influences. *Nature Reviews. Genetics*, 12(3): 167–78, 2011.

[18] M. P. Epstein, C. D. Veal, R. C. Trembath, J. N. Barker, C. Li, and G. A. Satten. Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet*, 76(4):592–608, 2005.

[19] J. Eu-Ahsunthornwattana, E. N. Miller, M. Fakiola, S. M. Jeronimo, J. M. Blackwell, and H. J. Cordell. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet*, 10(7):e1004445, 2014.

[20] G. Ferone, H. A. Thomason, D. Antonini, L. De Rosa, B. Hu, M. Gemei, H. Zhou, R. Ambrosio, D. P. Rice, D. Acampora, H. van Bokhoven, L. Del Vecchio, M. I. Koster, G. Tadini, B. Spencer-Dene, M. Dixon, J. Dixon, and C. Missero. Mutant p63 causes defective expansion of ectodermal progenitor cells and impaired fgf signalling in aec syndrome. *EMBO Mol Med*, 4(3):192–205, 2012.

[21] P. Fogh-Andersen. *Inheritance of Harelip and Cleft Palate.* Munksgaard, Copenhagen, 1942.

[22] A. Fomenkov, Y. P. Huang, O. Topaloglu, A. Brechman, M. Osada, T. Fomenkova, E. Yuriditsky, B. Trink, D. Sidransky, and E. Ratovitski. P63 alpha mutations lead to aberrant splicing of keratinocyte growth factor receptor in the hay-wells syndrome. *J Biol Chem*, 278(26):23906–14, 2003.

[23] F. C. Fraser. Thoughts on the etiology of clefts of the palate and lip. *Acta Genetica et Statistica Medica*, 5(4):358–69, 1955.

[24] S. F. Grant, K. Wang, H. Zhang, W. Glaberson, K. Annaiah, C. E. Kim, J. P. Bradfield, J. T. Glessner, K. A. Thomas, M. Garris, E. C. Frackelton, F. G. Otieno, R. M. Chiavacci, H. D. Nah, R. E. Kirschner, and H. Hakonarson. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *Journal of Pediatrics*, 155(6):909–13, 2009.

[25] D. Grosen, C. Chevrier, A. Skytthe, C. Bille, K. Molsted, A. Sivertsen, J. Murray, and K. Christensen. 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. *Journal of Medical Genetics*, 47(3):162–168, 2010.

[26] K. K. Gundlach and C. Maus. Epidemiological studies on the frequency of clefts in europe and world-wide. *Journal of Cranio-Maxillo-Facial Surgery*, 34 Suppl 2:1–2, 2006.

[27] D. B. Hancock and W. K. Scott. Population-based case-control association studies. *Curr Protoc Hum Genet*, Chapter 1:Unit1.17, 2012.

[28] Z. He, B. J. O'Roak, J. D. Smith, G. Wang, S. Hooker, R. L. Santos-Cortez, B. Li, M. Kan, N. Krumm, D. A. Nickerson, J. Shendure, E. E. Eichler, and S. M. Leal. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am J Hum Genet*, 94(1):33–46, 2014.

[29] C. Infante-Rivard, L. Mirea, and S. B. Bull. Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am J Epidemiol*, 170(5):657–64, 2009.

[30] Z. Jia, E. J. Leslie, M. E. Cooper, A. Butali, J. Standley, J. Rigdon, S. Suzuki, A. Gongorjav, T. E. Shonkhuuz, N. Natsume, B. Shi, M. L. Marazita, and J. C. Murray. Replication of 13q31.1 association in nonsyndromic cleft lip with cleft palate in europeans. *Am J Med Genet A*, 167(5):1054–60, 2015.

[31] R. Jiang, J. O. Bush, and A. C. Lidral. Development of the upper lip: morphogenetic and molecular mechanisms. *Dev Dyn*, 235(5):1152–66, 2006.

[32] J. L. Jones, J. W. Canady, J. T. Brookes, G. L. Wehby, J. L'Heureux, B. C. Schutte, J. C. Murray, and M. Dunnwald. Wound complications after cleft repair in children with van der woude syndrome. *Journal of Craniofacial Surgery*, 21(5):1350–3, 2010.

[33] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–54, 2010.

[34] G. R. Kazeem and M. Farrall. Integrating case-control and tdt studies. *Ann Hum Genet*, 69(Pt 3):329–35, 2005.

[35] A. Korte and A. Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9:29, 2013.

[36] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265(5181): 2037–48, 1994.

[37] P. H. Lee, S. E. Bergen, R. H. Perlis, P. F. Sullivan, P. Sklar, J. W. Smoller, and S. M. Purcell. Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework. *Hum Hered*, 72(1):10–20, 2011.

[38] E. Leslie and J. Murray. Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clin Genet*, pages n/a–n/a, 2012.

[39] E. J. Leslie and M. L. Marazita. Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet*, 163C(4):246–58, 2013.

[40] E. J. Leslie, M. A. Taub, H. Liu, K. M. Steinberg, D. C. Koboldt, Q. Zhang, J. C. Carlson, J. B. Hetmanski, H. Wang, D. E. Larson, R. S. Fulton, Y. A. Kousa, W. D. Fakhouri, A. Naji, I. Ruczinski, F. Begum, M. M. Parker, T. Busch, J. Standley, J. Rigdon, J. T. Hecht, A. F. Scott, G. L. Wehby, K. Christensen, A. E. Czeizel, F. W. Deleyiannis, B. C. Schutte, R. K. Wilson, R. A. Cornell, A. C. Lidral, G. M. Weinstock, T. H. Beaty, M. L. Marazita, and J. C. Murray. 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.

[41] E. J. Leslie, J. C. Carlson, J. R. Shaffer, E. Feingold, G. Wehby, C. A. Laurie, D. Jain, C. C. Laurie, K. F. Doheny, T. McHenry, J. Resick, C. Sanchez, J. Jacobs, B. Emanuele, A. R. Vieira, K. Neiswanger, A. C. Lidral, L. C. Valencia-Ramirez, A. M. Lopez-Palacio, D. R. Valencia, M. Arcos-Burgos, A. E. Czeizel, L. L. Field, C. D. Padilla, E. M. Cutiongco-de la Paz, F. Deleyiannis, K. Christensen, R. G. Munger, R. T. Lie, A. Wilcox, P. A. Romitti, E. E. Castilla, J. C. Mereb, F. A. Poletta, I. M. Orioli, F. M. Carvalho, J. T. Hecht, S. H. Blanton, C. J. Buxo, A. Butali, P. A. Mossey, W. L. Adeyemo, O. James, R. O. Braimah, B. S. Aregbesola, M. A. Eshete, F. Abate, M. Koruyucu, F. Seymen, L. Ma, J. E. de Salamanca, S. M. Weinberg, L. Moreno, J. C. Murray, and M. L. Marazita. 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.

[42] E. J. Leslie, H. Liu, J. C. Carlson, J. R. Shaffer, E. Feingold, G. Wehby, C. A. Laurie, D. Jain, C. C. Laurie, K. F. Doheny, T. McHenry, J. Resick, C. Sanchez, J. Jacobs, B. Emanuele, A. R. Vieira, K. Neiswanger, J. Standley, A. E. Czeizel, F. Deleyiannis, K. Christensen, R. G. Munger, R. T. Lie, A. Wilcox, P. A. Romitti, L. L. Field, C. D. Padilla, E. M. Cutiongco-de la Paz, A. C. Lidral, L. C. Valencia-Ramirez, A. M. Lopez-Palacio, D. R. Valencia, M. Arcos-Burgos, E. E. Castilla, J. C. Mereb, F. A. Poletta, I. M. Orioli, F. M. Carvalho, J. T. Hecht, S. H. Blanton, C. J. Buxo, A. Butali, P. A. Mossey, W. L. Adeyemo, O. James, R. O. Braimah, B. S. Aregbesola, M. A. Eshete, M. Deribew, M. Koruyucu, F. Seymen, L. Ma, J. E. de Salamanca, S. M. Weinberg, L. Moreno, R. A. Cornell, J. C. Murray, and M. L. Marazita. A genome-wide association study of nonsyndromic cleft palate identifies an etiologic missense variant in grhl3. *Am J Hum Genet*, 98(4):744–54, 2016.

[43] E. J. Leslie, J. C. Carlson, J. R. Shaffer, A. Butali, C. J. Buxo, E. E. Castilla, K. Christensen, F. W. Deleyiannis, L. Leigh Field, J. T. Hecht, L. Moreno, I. M. Orioli, C. Padilla, A. R. Vieira, G. L. Wehby, E. Feingold, S. M. Weinberg, J. C. Murray, T. H. Beaty, and M. L. Marazita. Genome-wide meta-analyses of nonsyndromic oro-

facial clefts identify novel associations between foxe1 and all orofacial clefts, and tp63 and cleft lip with or without cleft palate. *Hum Genet*, 136(3):275–286, 2017.

[44] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–21, 2008.

[45] D. Lindley. Statistical inference concerning hardy-weinberg equilibirum, 1987.

[46] K. U. Ludwig, E. Mangold, S. Herms, S. Nowak, H. Reutter, A. Paul, J. Becker, R. Herberz, T. AlChawa, E. Nasser, A. C. Bohmer, M. Mattheisen, M. A. Alblas, S. Barth, N. Kluck, C. Lauster, B. Braumann, R. H. Reich, A. Hemprich, S. Potzsch, B. Blaumeiser, N. Daratsianos, T. Kreusch, J. C. Murray, M. L. Marazita, I. Ruczinski, A. F. Scott, T. H. Beaty, F. J. Kramer, T. F. Wienker, R. P. Steegers-Theunissen, M. Rubini, P. A. Mossey, P. Hoffmann, C. Lange, S. Cichon, P. Propping, M. Knapp, and M. M. Nothen. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet*, 44(9):968–71, 2012.

[47] K. U. Ludwig, S. T. Ahmed, A. C. Bohmer, N. B. Sangani, S. Varghese, J. Klamt, H. Schuenke, P. Gultepe, A. Hofmann, M. Rubini, K. A. Aldhorae, R. P. Steegers-Theunissen, A. Rojas-Martinez, R. Reiter, G. Borck, M. Knapp, M. Nakatomi, D. Graf, E. Mangold, and H. Peters. Meta-analysis reveals genome-wide significance at 15q13 for nonsyndromic clefting of both the lip and the palate, and functional analyses implicate grem1 as a plausible causative gene. *PLoS Genet*, 12(3):e1005914, 2016.

[48] S. Luo, B. Mukherjee, J. Chen, and N. Chatterjee. Shrinkage estimation for robust and efficient screening of single-snp association from case-control genome-wide association studies. *Genet Epidemiol*, 33(8):740–50, 2009.

[49] E. Mangold, K. U. Ludwig, S. Birnbaum, C. Baluardo, M. Ferrian, S. Herms, H. Reutter, N. A. de Assis, T. A. Chawa, M. Mattheisen, M. Steffens, S. Barth, N. Kluck, A. Paul, J. Becker, C. Lauster, G. Schmidt, B. Braumann, M. Scheer, R. H. Reich, A. Hemprich, S. Potzsch, B. Blaumeiser, S. Moebus, M. Krawczak, S. Schreiber, T. Meitinger, H. E. Wichmann, R. P. Steegers-Theunissen, F. J. Kramer, S. Cichon, P. Propping, T. F. Wienker, M. Knapp, M. Rubini, P. A. Mossey, P. Hoffmann, and M. M. Nothen. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics*, 42(1):24–6, 2010.

[50] E. Mangold, A. C. Bohmer, N. Ishorst, A. K. Hoebel, P. Gultepe, H. Schuenke, J. Klamt, A. Hofmann, L. Golz, R. Raff, P. Tessmann, S. Nowak, H. Reutter, A. Hemprich, T. Kreusch, F. J. Kramer, B. Braumann, R. Reich, G. Schmidt, A. Jager, R. Reiter, S. Brosch, J. Stavusis, M. Ishida, R. Seselgyte, G. E. Moore, M. M. Nothen, G. Borck, K. A. Aldhorae, B. Lace, P. Stanier, M. Knapp, and K. U. Ludwig. Sequencing the grhl3 coding region reveals rare truncating mutations and a common susceptibility variant for nonsyndromic cleft palate. *Am J Hum Genet*, 98(4):755–62, 2016.

[51] M. L. Marazita. The evolution of human genetic studies of cleft lip and cleft palate. *Annu Rev Genomics Hum Genet*, 13:263–83, 2012.

[52] M. L. Marazita, A. C. Lidral, J. C. Murray, L. L. Field, B. S. Maher, T. Goldstein McHenry, M. E. Cooper, M. Govil, S. Daack-Hirsch, B. Riley, A. Jugessur, T. Felix, L. Morene, M. A. Mansilla, A. R. Vieira, K. Doheny, E. Pugh, C. Valencia-Ramirez, and M. Arcos-Burgos. Genome scan, fine-mapping, and candidate gene analysis of nonsyndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Human Heredity*, 68(3):151–70, 2009.

[53] P. A. Mossey and J. Little. Epidemiology of oral clefts: an international perspective. In D. Wyszynski, editor, *Cleft lip and palate: from origin to treatment.*, pages 127–158. Oxford University Press, Oxford, 2002.

[54] B. Mukherjee and N. Chatterjee. Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–94, 2008.

[55] T. Murray, M. A. Taub, I. Ruczinski, A. F. Scott, J. B. Hetmanski, H. Schwender, P. Patel, T. X. Zhang, R. G. Munger, A. J. Wilcox, X. Ye, H. Wang, T. Wu, Y. H. Wu-Chou, B. Shi, S. H. Jee, S. Chong, V. Yeow, J. C. Murray, M. L. Marazita, and T. H. Beaty. Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between asians and europeans. *Genetic Epidemiology*, 36(4):392–9, 2012.

[56] N. J. Nagelkerke, B. Hoebee, P. Teunis, and T. G. Kimman. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet*, 12(11):964–70, 2004.

[57] K. Neiswanger, S. M. Weinberg, C. R. Rogers, C. A. Brandon, M. E. Cooper, K. M. Bardi, F. W. Deleyiannis, J. M. Resick, A. Bowen, M. P. Mooney, J. E. de Salamanca, B. Gonzalez, B. S. Maher, R. A. Martin, and M. L. Marazita. Orbicularis oris muscle defects as an expanded phenotypic feature in nonsyndromic cleft lip with or without cleft palate. *Am J Med Genet A*, 143a(11):1143–9, 2007.

[58] K. Neiswanger, K. Walker, C. M. Klotz, M. E. Cooper, K. M. Bardi, C. A. Brandon, S. M. Weinberg, A. R. Vieira, R. A. Martin, A. E. Czeizel, E. E. Castilla, F. A. Poletta, and M. L. Marazita. Whorl patterns on the lower lip are associated with nonsyndromic cleft lip with or without cleft palate. *Am J Med Genet A*, 149a(12):2673–9, 2009.

[59] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*, 11(20):2417–23, 2002.

[60] R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–7, 2010.

[61] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3): 559–75, 2007.

[62] A. Rada-Iglesias, R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut, and J. Wysocka. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell*, 11(5):633–48, 2012.

[63] F. Rahimov, M. L. Marazita, A. Visel, M. E. Cooper, M. J. Hitchler, M. Rubini, F. E. Domann, M. Govil, K. Christensen, C. Bille, M. Melbye, A. Jugessur, R. T. Lie, A. J. Wilcox, D. R. Fitzpatrick, E. D. Green, P. A. Mossey, J. Little, R. P. Steegers-Theunissen, L. A. Pennacchio, B. C. Schutte, and J. C. Murray. Disruption of an ap-2alpha binding site in an irf6 enhancer is associated with cleft lip. *Nature Genetics*, 40(11):1341–7, 2008.

[64] F. Rahimov, A. Jugessur, and J. C. Murray. Genetics of nonsyndromic orofacial clefts. *Cleft Palate Craniofac J*, 49(1):73–91, 2012.

[65] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 1996.

[66] N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405 (6788):847–56, 2000.

[67] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. Encode data in the ucsc genome browser: year 5 update. *Nucleic Acids Res*, 41(Database issue):D56–63, 2013.

[68] D. J. Schaid and S. S. Sommer. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*, 53(5):1114–26, 1993.

[69] N. Schenker and J. F. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

[70] H. Schwender, Q. Li, C. Neumann, M. A. Taub, S. G. Younkin, P. Berger, R. B. Scharpf, T. H. Beaty, and I. Ruczinski. Detecting disease variants in case-parent trio studies using the bioconductor software package trio. *Genet Epidemiol*, 38(6):516–22, 2014.

[71] R. S. Spielman, 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*, 52(3):506–16, 1993.

[72] L. A. Stefanski and D. D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[73] G. R. Svishcheva, N. M. Belonogova, and T. I. Axenovich. Ffbskat: fast family-based sequence kernel association test. *PLoS One*, 9(6):e99407, 2014.

[74] D. Thomas. *Statistical methods in genetic epidemiology*. Oxford University Press, New York, 2004.

[75] H. A. Thomason and M. J. Dixon. Craniofacial defects and cleft lip and palate. In *ENCYCLOPEDIA OF LIFE SCIENCES*. John Wiley and Sons, Ltd, 2009.

[76] A. P. Vanderas. Incidence of cleft lip, cleft palate, and cleft lip and palate among races: a review. *Cleft Palate J*, 24(3):216–25, 1987.

[77] A. Visel, E. M. Rubin, and L. A. Pennacchio. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, 2009.

[78] Y. Wang, Y. Sun, Y. Huang, Y. Pan, Z. Jia, L. Ma, L. Ma, F. Lan, Y. Zhou, J. Shi, X. Yang, L. Zhang, H. Jiang, M. Jiang, A. Yin, J. Cheng, L. Wang, Y. Yang, and B. Shi. Association study between van der woude syndrome causative gene grhl3 and nonsyndromic cleft lip with or without cleft palate in a chinese cohort. *Gene*, 588(1): 69–73, 2016.

[79] G. L. Wehby and C. H. Cassell. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis*, 16(1):3–10, 2010.

[80] S. M. Weinberg, K. Neiswanger, R. A. Martin, M. P. Mooney, A. A. Kane, S. L. Wenger, J. Losee, F. Deleyiannis, L. Ma, J. E. De Salamanca, A. E. Czeizel, and M. L. Marazita. The pittsburgh oral-facial cleft study: expanding the cleft phenotype. background and justification. *Cleft Palate Craniofac J*, 43(1):7–20, 2006.

[81] S. M. Weinberg, S. D. Naidoo, K. M. Bardi, C. A. Brandon, K. Neiswanger, J. M. Resick, R. A. Martin, and M. L. Marazita. Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. *Orthod Craniofac Res*, 12(4):271–81, 2009.

[82] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*, 89(1):82–93, 2011.

[83] X. Zhan, Y. Hu, B. Li, G. R. Abecasis, and D. J. Liu. Rvtests: An efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics*, 2016.

[84] A. Ziegler and I. Knig. *A Statistical Approach to Genetic Epidemiology*. Wiley-VCH Verlag GmbH and Co. KGaA, Weinheim, 2nd edition edition, 2010.