Author:
Howe, Laurence

Title:
Exploring the aetiology and phenotypic consequences of non-syndromic cleft lip/palate using polygenic risk scoring and Mendelian randomization

Exploring the aetiology and phenotypic consequences of non-syndromic cleft lip/palate using polygenic risk scoring and Mendelian randomization

Laurence James Ming San Howe

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor or Philosophy in the Bristol Medical School

MRC Integrative Epidemiology Unit

Department of Population Health Sciences

Bristol Medical School

University of Bristol

Bristol, UK

May 11th, 2018

Word Count: 40,414

Abstract

Non-syndromic cleft lip/palate (nsCL/P) is a congenital birth defect characterised by cleft(s) of the upper lip with or without a cleft of the palate. The aetiology of nsCL/P is complex with both genetic and environmental risk factors. In this thesis, I applied Mendelian randomization (MR) and polygenic risk scoring (PRS) to explore the aetiology of nsCL/P and possible consequences of the phenotype.

In **Chapter 3**, strong evidence was found for nsCL/P having a highly polygenic architecture with a substantial SNP heritability suggesting that PRS are likely to be effective genetic proxies for nsCL/P.

In **Chapter 4**, three putative loci were identified where the effect of nsCL/P genetic risk variants on disease liability may be mediated by DNA methylation, although conclusions are limited by possible tissue specific effects.

In **Chapter 5**, nsCL/P genetic risk variants were shown to have a consistent additive effect on philtrum width in the general population suggesting that liability to nsCL/P causes decreased philtrum width which supports a polygenic threshold model of inheritance for nsCL/P.

In **Chapter 6**, nsCL/P genetic variants were shown to not be strongly associated with adverse developmental outcomes that are common in nsCL/P cases. These findings suggest that the adverse outcomes tested are unlikely to be related to underlying liability to nsCL/P.

In **Chapter 7**, nsCL/P PRS were found to predict increased risk of oral cavity/oropharyngeal cancer (OC/OPC). Follow-up analyses suggested the relationship was non-causal and that nsCL/P and OC/OPC likely share risk factors, possibly environmental risk factors or biological processes.

Acknowledgements

I greatly appreciate the contributions of the following individuals, groups and organisations to the work undertaken in this doctoral thesis:

● My PhD supervisors: Dr Sarah Lewis, Dr Gibran Hemani, Dr Beate St. Pourcain and Professor George Davey Smith for their support, guidance and imparted wisdom.

● Members of the Cleft Collective, in particular Dr Gemma Sharp, Dr Evie Stergiakouli and Professor Jonathan Sandy for support and guidance.

● The families and individuals participating in or involved in the data collection for the many different studies that have contributed data to analyses in this thesis, which include: the Avon Longitudinal Study of Parents and Children, the International Cleft Consortium via dbGAP, the UK Biobank, the UK Biobank, GOYA and The Cleft Collective.

● The Medical Research Council (MRC) and Integrative Epidemiology Unit (IEU) for the funding of my 4-year studentship.

● My external collaborators, from whom I have learned a great deal and whom have contributed immensely, including: Seth and the Pittsburgh group, Stephen and the Cardiff group, Elisabeth and the Bonn group, Ellen and the GOYA team, Maria-Rita and her team, and Paul Brennan and the genetic epidemiology group at IARC.

● A large number of colleagues in the MRC-IEU at the University of Bristol, for immeasurable assistance and advice.

● My parents for their support throughout the four years.

● My friends in Bristol for both great social experiences to keep me sane and the thousands of hours spent discussing epidemiology and genetics outside of work. Stimulating conversations with JY, TB, MACV, RA, JB and JE were particularly plentiful.

● The many other academics within the scientific community who inspired me directly through conversation or indirectly through reading their thought-provoking work.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Expect where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: Date:

Table of Contents

List of Tables

List of Figures

Commonly used abbreviations

3DFN: 3D Facial Norms Study

ALSPAC: Avon Longitudinal Study of Parents and Children

ARIES: Accessible Resource for Integrated Epigenomic Studies

ASD: Autism spectrum disorders

AVENGEME: Additive Variance Explained and Number of Genetic Effects Method of Estimation

CEU: Refers to individuals of northern European descent from Utah in the 1000 Genomes

CL/P: Cleft lip with/without cleft palate

CP/L: Cleft palate with/without cleft lip

CL: Cleft lip

CLO: Cleft lip only

CLP: Cleft lip AND cleft palate

CP: Cleft palate

CPO: Cleft palate only

DZ: Dizygotic (twins)

eQTL: Gene-expression quantitative trait loci

EWAS: Epigenome-wide Association Study

GBR: Refers to individuals of European descent from Great Britain in the 1000 Genomes

GCTA: Genome-wide Complex Trait Analysis

GOYA: Genetics of Overweight Young Adults

GREML: Genetic relationship matrix restricted maximum likelihood

GRM: Genetic relationship matrix

GTex: The Genotype-tissue expression project

GWAS: Genome-wide Association Study

HWE: Hardy-Weinberg Equilibrium

IBD: Identical by descent

IBS: Identical by state

ICC: International Cleft Consortium

IV: Instrumental Variable

IVW: Inverse Variance Weighted

JLIM: Joint likelihood mapping (an R package)

LD: Linkage Disequilibrium

LDAK: Linkage disequilibrium adjusted kinships

LMM: Linear mixed models

MAF: Minor allele frequency

MEP: Middle ear pressure

mQTL: Methylation quantitative trait loci

MR: Mendelian randomization

MZ: Monozygotic (twins)

NNC: NESDA NTR Conditional eQTL Catalog

nsCL/P: Non-syndromic cleft lip with/without cleft palate

OC/OPC: Oral cavity, oropharyngeal or hypopharyngeal cancer

OC: Oral cavity cancer

OFC(s): Orofacial cleft(s)

OME: Otitis media with effusion

OPC: Oropharyngeal cancer

OR: Odds ratio

PCA: Principal components analysis

PRS: Polygenic risk scoring/scores/score

PSD: Persistent Speech Disorder

PTDT: Polygenic Transmission Disequilibrium Test

QC: Quality control

RCT: Randomised controlled trial

REML: Restricted estimated maximum likelihood

S.D.: Standard deviation(s)

SNP: Single nucleotide polymorphism

T2D: Type two diabetes

TDT: Transmission Disequilibrium Test

95% C.I.: 95% Confidence Interval

Research Outputs

Preprints

Howe, Laurence J., et al. "DNA methylation mediates genetic liability to non-syndromic cleft lip/palate." *bioRxiv* (2018): 256842. (Chapter 4)

Howe, Laurence J., et al. "Investigating the shared genetics of non-syndromic cleft lip/palate and facial morphology." *bioRxiv*(2018): 255901. (Chapter 5)

Chapter 1: Introduction

The focus of this doctoral thesis is non-syndromic cleft lip/palate (nsCL/P), a subtype of orofacial clefts (OFCs). Previous research has established that nsCL/P has a complex aetiology with both genetic and environmental risk factors. Although nsCL/P does not typically present with major anomalies independent from the facial cleft structure, there is growing evidence for phenotypic differences between nsCL/P cases and the general population, such as lower educational attainment amongst nsCL/P cases.

Despite recent advances in nsCL/P research there remains limited or partial understanding in specific areas, such as; the characteristics of nsCL/P-related genetic variation (genetic architecture), the aetiological role of epigenetic processes for nsCL/P, the nature of the shared genetics between nsCL/P and facial morphology, and the mechanisms underlying the increased incidence of developmental and disease outcomes in nsCL/P cases. In this chapter, I introduce the nsCL/P phenotype and the research questions of interest before discussing the advantages of applying Mendelian randomization (MR) and polygenic risk scoring (PRS) approaches to explore correlation and causality to increase understanding of the phenotype.

1.1 Non-syndromic cleft lip/palate

In this section, I give an overview of what is currently known about nsCL/P, including information relevant to the aetiology, phenotypic consequences and public health burden.

1.1.1 Facial development and orofacial clefts

OFCs refer to a subtype of facial clefts, primarily affecting the upper lip and the palate [1]. The facial region begins to form in the fourth embryonic week; bilateral cellular structures called pharyngeal or branchial arches create processes that coordinate facial development. The upper lip and upper jaw are formed by the fusion of the medial nasal and maxillary processes, the lower jaw is formed by the mandibular process while the philtrum is formed by the fusion of the two medial nasal processes [2,3]. The palate begins to form by the sixth week and is formed by several palatine processes. A cellular mass developing between the surfaces of the maxilla forms the primary palate while the second palate is formed by the remaining hard and soft palate [4]. Even minor developmental anomalies may prevent process fusion, which results in the formation of a cleft at the boundary [3]. Therefore, clefts can form at any point where merging or fusion occurs during development; a complete cleft is characterised by the complete absence of merging or fusion while some merging or fusion is found in incomplete clefts [5].

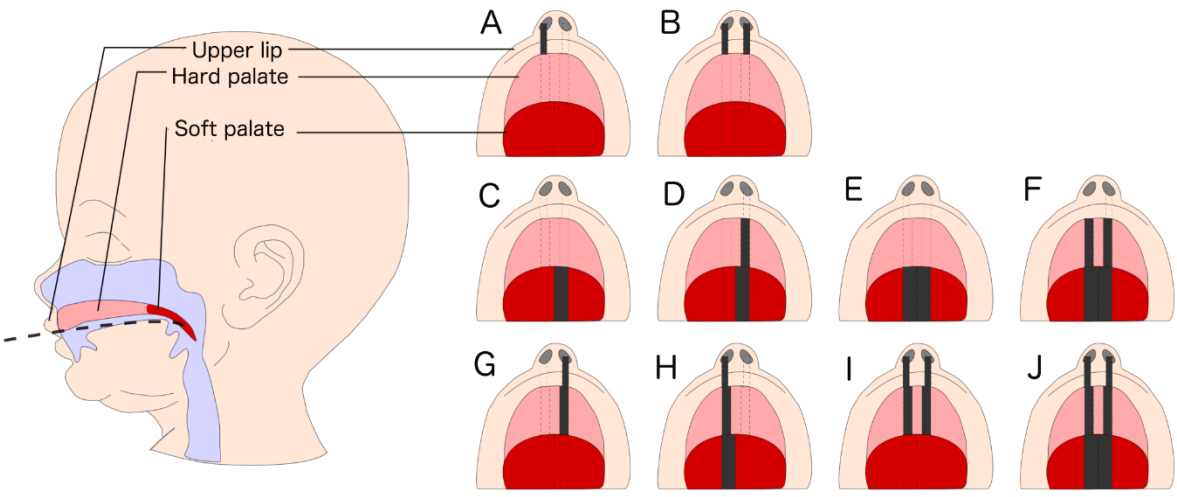
1.1.2 Incidence and subtypes of orofacial clefts

OFCs have an estimated incidence of around 1 in 700 individuals, with the incidence varying across different ethnic groups [1,6]. OFCs are often divided into the subtypes; cleft palate only (CPO) and cleft lip with/without cleft palate (CL/P). The subtypes are sexually dimorphic, with CL/P more common in males and CPO more frequent amongst females [1]. The argument for the division into CPO and CL/P is that embryology and genetic research suggest mechanistic differences between the development of clefts of the primary palate and lip with development of clefts affecting only the secondary palate [7]. However, there is also emerging evidence of genetic and epigenetic differences within the CL/P subtype, between cleft lip with

cleft palate (CLP) and cleft lip only (CLO), suggesting that these subtypes should also be considered separately when possible [8,9]. The different cleft subtypes can also occur unilaterally, i.e. affects one side of the lip or palate, or bilaterally, affects both sides of the lip or palate. **Figure 1** below shows 10 common OFC subtypes and demonstrates the heterogeneity within the different OFC sub-classifications (e.g. CL/P, CPO).

Figure 1: Cleft lip and palate – subtypes and classifications

Permission kindly granted for the use of the figure from Dr Gemma Sharp.



A – Unilateral cleft lip only (CLO or CL/P)

B – Bilateral cleft lip only (CLO or CL/P)

C – Unilateral cleft of the secondary palate only (CPO)

D – Unilateral clefts of the primary and secondary palates (CPO)

E – Bilateral cleft of the secondary palate (CPO)

F – Bilateral clefts of the primary and secondary palates (CPO)

G – Unilateral clefts of the lip and primary palate (CLP or CL/P)

H – Unilateral clefts of the lip and primary/secondary palates (CLP or CL/P)

I – Bilateral clefts of the lip and primary palate (CLP or CL/P)

J – Bilateral clefts of the lip and primary/secondary palates (CLP or CL/P)

1.1.3 Syndromic and non-syndromic orofacial clefts

OFCs can present as a symptom of a Mendelian disorder, traits with inheritance controlled by a single locus. One of the most common OFC syndromes is Van der Woude syndrome, which is caused by mutations in the *IRF6* gene [10]. There are over 400 genetic syndromes that may present with an OFC, among other symptoms [11], including Pierre-Robin sequence/syndrome, median facial dysplasia and Velocardiofacial syndrome [12]. Approximately 70% of OFCs are non-syndromic, presenting with the facial cleft structure but no other apparent developmental or physical abnormalities [7]. Diagnosis of a non-syndromic OFC is typically based on family history and the absence of other abnormalities. The focus of this thesis project is nsCL/P.

1.1.4 Aetiology of nsCL/P

While syndromic forms of cleft are caused by genetic variation in a single locus, nsCL/P has a more complex aetiology with multiple distinct genetic risk factors and several proposed environmental risk factors [13].

1.1.4.1 Genetic risk factors

Linkage analysis [14], an early genetic risk mapping method, was relatively successful for identifying regions relevant to syndromic OFCs [15,16] but was largely ineffective for nsCL/P. Multiple regions were found across different studies to co-segregate with nsCL/P [17-19] but the findings often did not replicate [20]. The lack of clear segregation within pedigrees demonstrated the complex genetic aetiology of nsCL/P.

Genome-Wide Association Studies (GWAS) were more successful in identifying risk loci for nsCL/P. Different study designs, including parent-offspring trios [21] and case-control designs [22,23] were used to test association between the

nsCL/P and the genome. Subsequent meta-analyses of GWAS results from independent data-sets, across distinct ethnic backgrounds, have been used to identify over 40 distinct genetic risk loci for nsCL/P [9,21-30]. The characteristics of nsCL/P genetic risk variants, such as allele frequency and effect sizes, have important implications for aetiological understanding of the phenotype as well as for study design. The genetic architecture of nsCL/P as well as relevant implications and methodologies (e.g. GWAS) are discussed and explored in more detail in **Chapter 3**.

Follow-up of nsCL/P risk loci demonstrated that the majority of risk regions contain only one common variant signal [31] and that a major nsCL/P risk locus may affect disease liability through gene expression pathways [32]. Epigenetic processes such as DNA methylation may play a role in the aetiology of nsCL/P [8,33-36] through gene expression. DNA methylation is of particular epidemiological interest to nsCL/P because it can be altered by environmental exposures. The aetiological relevance of DNA methylation to nsCL/P is the primary focus of **Chapter 4**.

1.1.4.2 Environmental risk factors

Environmental risk factors for nsCL/P are not as well-characterised as the genetic risk factors. The in-utero and maternal environment are the primary routes of exposure for a birth defect, and the effects of certain maternal environmental risk factors on embryology are well characterised. For example, maternal alcohol intake is known to have a teratogenic effect on the developing foetus [37,38] and maternal folic acid supplementation has been shown in randomised controlled trials (RCTs) to be protective against neural tube defects [39-41]. Other potential maternal risk factors during pregnancy, include; cigarette smoking, which is associated with increased risk of adverse outcomes such as low birth-weight [42-46], and obesity, which is associated with increased risk of congenital anomalies [47-49].

Previously studied maternal environmental risk factors for nsCL/P include; folic acid [50-66], obesity [48,67-69], smoking [70-76], alcohol consumption [74,77-80], cholesterol [81,82], low plasma zinc [83-85], vitamin B6 [54,86,87] and vitamin A [88]. In general, the evidence for association between proposed exposures and risk of nsCL/P is largely inconsistent, particularly for folic acid [57], but a recent meta-analysis found consistent evidence for an association between maternal adiposity and cleft palate with or without cleft lip (CP/L) [67]. It is also important to note that it is possible that the paternal environment may play an aetiological role via interaction effects with the mother or via sperm [89].

A major limitation of reviewing the epidemiological literature to infer aetiological relationships is that studies reporting null results may be less likely to be published, which can skew the literature evidence and affect meta-analyses [90]. A further limitation is that observational epidemiological studies may be susceptible to confounding (discussed in the next section). Reliably identifying maternal risk factors for nsCL/P could have important implications for prevention and also improve biological understanding. An MR framework, which will be discussed in the next section, has the potential to accurately investigate possible maternal risk factors for nsCL/P. However, an investigation of maternal risk factors for nsCL/P is greatly limited by the available data sources (data sources are described in **Chapter 2**), which lack sufficient sample sizes and suitable control groups. This is discussed in more detail in **Chapter 8**.

1.1.5 Phenotypic consequences

Typically, the treatment for an OFC involves surgical corrections to repair the lip and/or palate of the affected child, at a young age. However, post-surgery, there is substantial evidence that individuals born with an OFC differ phenotypically from

the general population. Examples of phenotypic differences between nsCL/P cases and controls that will be discussed in more detail in this section, include: facial morphological differences distinct from the OFC, dental anomalies and difficulties with speech [1]. More concerning is the evidence suggesting that individuals with nsCL/P may have a reduced quality of life and higher mortality up until the age of 55 [13,91,92].

1.1.5.1 Facial morphology

Syndromes including an OFC often present with additional craniofacial anomalies, but nsCL/P does not usually present with major facial anomalies distinct from the OFC. However, there are several lines of evidence suggesting a possible relationship between the genetics of nsCL/P and facial morphology. Firstly, the genetic overlap between OFC syndromes and nsCL/P suggests the potential for some overlap in phenotypic presentation, independent of the facial cleft structure [93,94], secondly, the identification of pleiotropic genes implicated in both nsCL/P and facial variation [95], and thirdly, the observation of sub-clinical differences in facial morphology in individuals with nsCL/P and their unaffected relatives [93,96-98]. Facial development and the formation of an OFC are largely synchronous, so improved understanding of shared genetics may have important aetiological implications. The relationship between genetic variants associated with nsCL/P and facial morphology is explored and discussed in more detail in **Chapter 5**.

1.1.5.2 Adverse developmental outcomes

Adverse outcomes with evidence for increased incidence in nsCL/P cases, include; impaired growth [99-101], dental anomalies [102-109], auditory problems [110-113], speech impediments [114-116], behavioural problems [117-120] and lower educational attainment [121]. The increased risk of these adverse outcomes means that individuals

with an OFC may have reduced quality of life [13,91]. It is currently unclear which outcomes are related to an individual's underlying liability to nsCL/P (the liability concept is discussed in more detail in the next section) and which outcomes are caused by the physical presence of an OFC and/or the corrective surgery. Discerning which of the two possibilities is more likely may have important implications for treatment and follow-up of nsCL/P cases. In **Chapter 6**, the association between genetic risk variants for nsCL/P and adverse developmental outcomes is investigated.

1.1.5.3 Increased risk of cancer?

Previous research has suggested that congenital anomalies and cancer may have shared aetiology [122]. Specifically for nsCL/P, there is largely inconsistent evidence for increased risk of cancer amongst cases and their unaffected family members [122-126] but several genes have been implicated in the aetiology of both phenotypes [127-129].

Cancers of the oral cavity and pharynx (OC/OPC) are strong candidates for having shared genetic aetiology with nsCL/P and this potential relationship has not been previously investigated. Shared genetics between nsCL/P and OC/OPC would suggest that nsCL/P cases may have increased risk of OC/OPC and that the two phenotypes may share common biological pathways or environmental risk factors. In **Chapter 7**, I investigate the possibility of shared genetic aetiology between OC/OPC and nsCL/P.

1.2 Epidemiology, the genome, polygenic risk scores & Mendelian randomization

The research questions of this doctoral thesis are epidemiological in nature, involving exploring the causes and consequences of nsCL/P at a population level.

Here I discuss different epidemiological techniques and explain the relevance of the genome in an epidemiological context. MR and PRS are recurring themes throughout this thesis, both are utilised in the majority of results chapters and so are described in detail.

1.2.1 Correlation & Causation

Epidemiology refers to the study of the effects and causes of disease. One of the most important public health applications of epidemiology is to inform preventative interventions, which aim to reduce the incidence of a disease in a population or a sub-sample of a population. Interventions may take a number of forms, including medical advice, drug prescriptions or government policy. For example, based on evidence that red meat may be carcinogenic, the World Health Organisation recommended reduced consumption of red meat [130].

The premise of a preventative intervention is that on a population level, the intervention will cause a beneficial change, and so prevention is heavily reliant on genuine causal relationships. Correlation does not necessarily imply causation, distinct traits can be phenotypically correlated, i.e. two traits can often occur together in the same individuals, but this does not mean that modifying one trait would directly impact the other. One possibility is that two traits are correlated because of shared causes of the traits; for example, yellow fingers and lung cancer are correlated because they are both caused by cigarette smoking. Alternatively, this correlation could be because one of the traits is causally linked to the other, as in the genuine causal relationship between tobacco smoking and risk of lung cancer [131].

1.2.2 Randomised controlled trials

The gold standard for testing causality is the RCT. RCTs aim to enact a controlled experiment by randomising individuals to a treatment or placebo group

with study participants and personnel blinded to treatment group allocation. Study participants are then followed up over time, to determine the effects of the treatment on predetermined outcomes. When designed appropriately and well-powered, RCTs provide reliable causal estimates but trials cannot be used to test every potential causal relationship because they are expensive, time consuming and often unethical.

1.2.3 Observational epidemiological studies

One epidemiological tool that can help to identify causal relationships is observational epidemiology. Observational epidemiology studies involve observing and collecting data from a sub-sample of a population and extrapolating the results in order to make population-wide level inferences. Case-control, cross-sectional and prospective cohort studies are among the most frequently employed designs.

1.2.3.1 Observational study designs

Case-control studies are often retrospective; cases for a particular disease are recruited and compared to controls (individuals without the disease), purported causal attributes are compared between the two groups to identify factors that may contribute to disease incidence. Case-control studies are very useful for studying rare diseases or events but the potential for bias, especially sampling bias, is a major limitation. A cross-sectional study involves extracting data from a population at a single time point in order to determine the prevalence - the number of cases in a population - at a specific time-point. Cross-sectional studies can investigate multiple outcomes but because data are only extracted at a single time-point, cannot easily distinguish between cause and effect. Prospective cohort studies involve recruiting a representative sample from the population and following study participants up over time, to determine if they develop outcomes of interest. Cohort studies are an often-preferred study design because the recording of events over a long period of time

may alleviate the possibility of reverse causation, where the outcome of interest actually affects the exposure of interest. However, prospective cohort studies are often expensive, time-consuming and may be biased by confounders [132]. A notable successful prospective cohort study was the British Doctor's study which provided evidence supporting the causal link between lung cancer and smoking [131].

1.2.3.2 Limitations of observational studies

The primary limitation of observational studies is that they are not controlled experiments, like RCTs. To determine the causal effect of an exposure on a disease outcome, we ideally want to ensure that study participants are identical in every regard other than levels of the exposure. In general, across populations this does not hold; lifestyle and dietary factors are heavily interlinked. For example, individuals who have a diet low in selenium are likely to differ in other ways (e.g. in relation to other dietary factors or lifestyle), to individuals with a diet high in selenium. This leads to difficulties with inferring causality from observational studies because of the potential for confounding. Despite the best efforts to account for confounders in analysis, observational studies are still highly susceptible to unmeasured residual confounding. The difficulties of inferring causality from observational studies are highlighted by the discordance between observational studies and RCTs; for example, observational studies found increased beta-carotene intake was associated with decreased risk of cardiovascular disease but RCTs found weak evidence for a protective effect of beta-carotene supplementation [133]. The differing results may be explained by confounding or they may be explained by differences in the measured exposure (short term supplementation against circulating levels). Nevertheless, for informing potential interventions, the results of observational

analyses are often unreliable at robustly identifying causal risk factors, meaning that when RCTs are not possible, there are limited options for causal inference.

1.2.4 The use of genetic data in an epidemiological context: Polygenic risk scores and Mendelian randomization

An alternative approach to observational epidemiology is to use genetic data in an epidemiological context (genetic epidemiology) to explore correlation and causality. In this section, I will discuss two relevant methodologies, PRS and MR. PRS are scores consisting of multiple genetic variants associated with a trait that can be used for either risk prediction or for the detection of shared genetic factors influencing two distinct traits (which could suggest a possible causal relationship). MR is an instrumental variable (IV) approach using genetic variants associated with an exposure to formally examine the causal relationship between that exposure and an outcome.

1.2.4.1 Polygenic risk scores

PRS include genetic variants that are associated with the phenotype of interest, often identified in GWAS, and are used as genetic predictors or proxies of the phenotype of interest. PRS are determined for each individual using either an unweighted score, where the number of risk alleles are summed, or a weighted score where GWAS effect sizes are used to weight the contribution of each genetic variant to the score. PRS have the potential to be useful in risk prediction; a disease PRS derived from one study can be tested for prediction of risk of the same disease in an independent sample. Although PRS often explain a substantial proportion of phenotypic variation (~40% for height [134]), the use of PRS for risk prediction is still largely in its infancy. PRS can also be used to detect shared genetic risk factors between two distinct traits; if a PRS for a trait is associated with an observed

phenotype in an independent sample, then this suggests that the association is likely attributable to genetic factors [135]. A further application of PRS is to formally estimate genetic correlation, which will be detailed in **Chapter 3**.

PRS have had success as genetic proxies for complex traits where there are few individual genetic variants reaching genome-wide significance and where many common variants are thought to contribute to disease risk, i.e. highly polygenic traits [136]. By including a large number of genetic variants, PRS often have increased power to detect shared genetic aetiology compared to scores which include only genome-wide significant variants. The utility of using PRS to detect shared genetic aetiology between nsCL/P and other phenotypes, will be explored in **Chapter 3**.

However, a strong association between a PRS for trait A with trait B could imply either correlation or causality. The relationship could be non-causal; trait A and trait B have shared genetic aetiology. Alternatively, the relationship could be causal; if trait A has a causal effect on trait B, then the genetic risk factors for trait A will also influence trait B, or vice-versa. The likelihood and interpretation of the different possibilities is largely dependent on the specific traits of interest and relevant factors such as temporality and liability (which will be discussed later on in the chapter).

1.2.4.2 Premise of Mendelian randomization

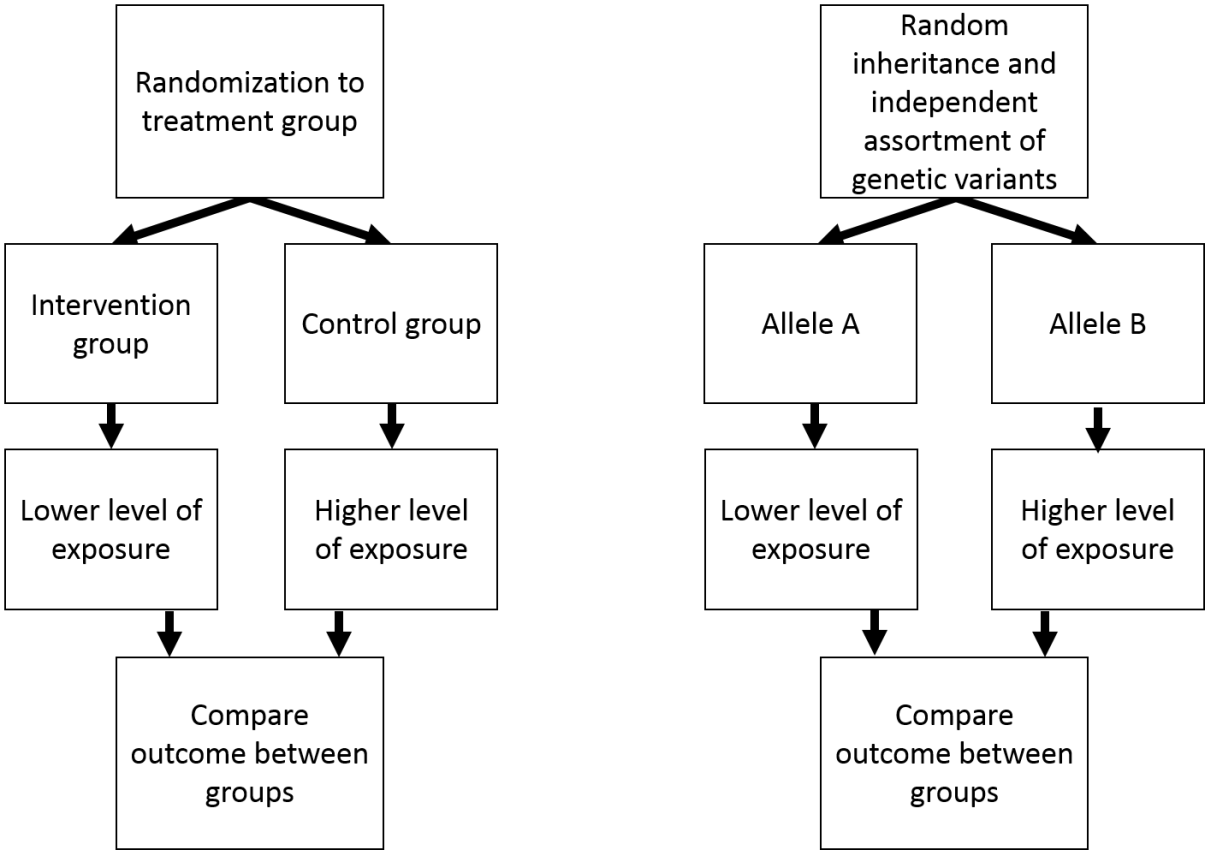
PRS can consist of hundreds or thousands of genetic variants and so, are often well-powered to detect associations. However, as described above, when there is an association between a PRS and an observed phenotype it can be difficult to distinguish between correlation and causality. Similar to PRS methods, MR analyses utilise genetic variants but under certain conditions, which will be detailed below, MR can be used to evaluate causality. Important differences between the two methods are first, that in MR analyses the inclusion of genetic variants is more stringent, and

second, that MR analyses often include sensitivity analyses to test assumptions required for inferring a causal relationship.

The primary strength of using an MR approach to evaluate causality over observational epidemiological methods is that genetic variation may be less prone to reverse-causation and confounding than the measured phenotype itself. Reverse-causation is largely ruled out because an individual's genotype is fixed from birth [137], while it follows from the first two laws of Mendelian genetics that genetic variants are unlikely to be associated with confounding factors on a population level. Mendel's first law of segregation states that allele pairs segregate during meiosis and randomly recombine at conception suggesting that each segregated allele has an equal chance of being present in a zygote. An extension of this law is that the probability of each segregated allele being present in a study participant can be assumed to be independent of environmental factors. While Mendel's second law of independent assortment states that alleles for different traits assort independently of one another, suggesting that the inheritance of a trait is independent of the inheritance of other traits [137,138].

If genetic variation leads to phenotypic differences which in turn affect disease risk, then the genetic variation should be related to the disease risk to the extent of the effect of the genetic variation on the phenotypic differences. The random inheritance of alleles and independent assortment of alleles relevant to an MR analysis offers parallels to the random allocation of treatment groups in an RCT, although there are several important distinctions [137] (**Figure 2**).

Figure 2: Comparison between a Randomised Controlled Trial (left) and a Mendelian randomization design (right) for estimating the causal effect of an exposure on an outcome



1.2.4.3 MR in a parent-offspring trio

MR can be illustrated by considering the pattern of inheritance within a parent-offspring trio; assuming one parent is heterozygous at a locus and the other is homozygous, then the offspring has a 50% chance of inheriting each of the alleles from the heterozygous parent. Therefore, if the genotype at this locus is causally related to a disease, we would expect a higher proportion of disease cases to have been transmitted the disease risk allele from heterozygous parents than for controls. If an over-transmitted risk allele is known to be related to increased LDL cholesterol, this suggests that LDL cholesterol may be a risk factor for the disease. However, in practice, MR analyses are seldom applied to parent-offspring trios because of the

difficulties recruiting large cohorts of trios and the lower statistical power compared to case-control studies [137].

1.2.4.4 MR in association studies and one/two-sample MR

Therefore, MR analyses mostly utilise data from genetic association studies of individuals who are not closely related. MR in this context can be thought of as an approximation to a parent-offspring framework because the case-control design may be susceptible to potential sources of bias such as assortative mating and dynastic effects, which a parent-offspring design is robust to. Potential sources of bias in MR studies will be discussed later in the chapter.

MR analyses can be one-sample, where the association between the genetic instruments with the exposure and outcome are measured in the same individuals, or two-sample, where these associations are measured in non-overlapping samples. A major advantage of two-sample MR is that the exposure and outcome do not need to be measured in the same data-set. Instead, GWAS summary statistics for the exposure and outcome can be used to estimate the causal effect. Two-sample MR is particularly useful for the analyses in this thesis because a primary research question regards the causal effect of nsCL/P risk variants (derived from nsCL/P cases and controls) on outcomes in the general population. However, two-sample MR requires careful harmonisation of genetic instruments and the assumption that genetic instruments for the exposure are also valid instruments for the exposure in the outcome data-set [139].

1.2.4.5 Assumptions of Mendelian randomization

MR is not, however, a panacea to every epidemiological question. The method relies on several assumptions, which are required for causal inference. Firstly, variants must be robustly associated with the exposure (usually genome-wide

significant), secondly, the variants cannot influence the outcome through a pathway independent of the exposure and thirdly, the variants should not be associated with confounders of the relationship between the exposure and the outcome [140].

The first assumption requires the identification of genetic variation robustly associated with the exposure of interest in the population of study, often from GWAS, ideally replicated. Population stratification may lead to violations of the first assumption. The second and third assumptions may be violated for a number of different reasons, including; survival bias, pleiotropy, assortative mating, dynastic effects and canalisation [141].

1.2.4.6 Survival bias

Survival bias can affect the estimate of the gene-outcome relationships in a GWAS, which may violate the second and third assumptions. An illustrative example of where survival bias could potentially affect the gene-outcome estimate is in the context of tobacco smoking and Alzheimer's disease. A previous MR study, found that a genetic variant in *CHRNA5*, which is strongly associated with the quantity of cigarette smoking, is protective against Alzheimer's [142]. If smoking worsens the progression of Alzheimer's resulting in increased mortality amongst cases, then this would mean that cases with the smoking-related variant may be underrepresented in study populations. This underrepresentation could then induce a spurious inverse association between the smoking-related variant and Alzheimer's incidence.

1.2.4.7 Population stratification

Population stratification refers to genetic variants being spuriously associated with a trait when they are instead ancestral markers. If we assume that a disease, say type 2 diabetes (T2D), is much more prevalent in a particular ancestral group than another, then if the ancestral groups have defined genetic differences, many

ancestral markers will be associated with T2D for reasons unrelated to T2D [143]. For this reason, GWAS are generally performed within relatively homogeneous populations and methods such as principal components analysis (PCA) or linear mixed models (LMM) are further used to account for population differences between individuals.

1.2.4.8 Pleiotropy

Pleiotropy refers to a genetic variant being associated with multiple independent phenotypes. There are two types of pleiotropy; vertical or mediatory pleiotropy and horizontal pleiotropy (which can lead to balanced or directional pleiotropy in an MR context). Vertical or mediatory pleiotropy refers to the instance where a genetic variant affects the exposure, which in turn affects the outcome (i.e. the exposure and the outcome are on the same pathway and there is a causal relationship). Horizontal pleiotropy refers to the instance where a genetic variant affects the exposure and outcome through independent pathways. Balanced pleiotropy refers to the instance when genetic variants are horizontally pleiotropic, but across all genetic variants, the net bias in an MR analysis from pleiotropy is null [144], while directional pleiotropy refers to the instance when horizontal pleiotropy biases the causal estimate.

Wide-spread pleiotropy across the genome is thought to be the norm rather than the exception [145]. However, the effect of wide-spread pleiotropy on the validity of MR analyses is dependent on whether the pleiotropy is balanced or directional. A previous study found that for two highly polygenic traits, height and bone mineral density, there is no evidence of directional pleiotropy [146]. MR-Egger [147] can be used to formally test for unbalanced horizontal pleiotropy while comparisons between the primary effect estimate and different sensitivity analyses such as the weighted

median [148] and the weighted mode [149] can also be used to infer unbalanced horizontal pleiotropy.

1.2.4.9 Assortative mating

Assortative mating refers to the non-random mating of individuals according to a particular phenotype, characterised by a higher phenotypic correlation amongst spouse-pairs than in non-spouse pairs. For example, if we assume that a phenotype is influenced by both environmental and genetic factors, then assortative mating on the phenotype will result in spousal correlation for both environmental and genetic factors relating to the phenotype. In any resulting offspring, environmental and genetic factors relating to the phenotype will then be correlated. Genetic factors being associated with environmental factors, which are common confounders, violates the third IV assumption. There is strong evidence that many complex traits such as educational attainment and height influence mate selection and so care must be taken when using these phenotypes in an MR study [150-152].

1.2.4.10 Dynastic effects

In some instances, MR analyses may be biased by the genotype of the parents, otherwise known as dynastic effects. If we are trying to disentangle whether alcohol consumption reduces intelligence, then maternal genotype is a potential confounder. This is because if the mother has a variant influencing their alcohol consumption, then this will affect the genotype of the offspring, potentially the social environment for the offspring via alcohol consumption, potentially the drinking behaviour of their offspring through a shared environment and also potentially affect the IQ of the offspring because of the teratogenic impact of alcohol.

Effects of assortative mating and dynastic effects are alleviated when using parent-offspring designs, as the parental genotypes may be included in the model.

However, as discussed previously, in practice complete genotype data on large numbers of offspring and their parents is often unavailable and so, MR frequently utilises genetic instruments derived in genetic studies of unrelated individuals.

1.2.4.11 Canalisation

Canalisation refers to the degree to which phenotypic development is unaffected by genetic and environmental variation on a population level [153]. Canalisation may have important implications for MR because developmental compensation for the effect of a genetic variant may distort the estimate of the gene-outcome relationship [141].

1.2.4.12 *Applications of Mendelian randomization*

The conventional use of MR is for testing causal relationships between modifiable continuous exposures and disease outcomes (e.g. cholesterol or alcohol on cardiovascular disease [154,155]). Beyond determining the effects of continuous environmental exposures on disease outcomes, the principles of MR can be used more generally to infer the likely direction of causality. For example, in the context of DNA methylation. DNA methylation is an epigenetic process, which unlike the genome, varies across the life course. Therefore, DNA methylation can be a cause or consequence of a phenotype and so the principles of MR can be used to investigate temporal relationships; DNA methylation may be on the causal pathway to a phenotype or DNA methylation could be a consequence of a phenotype (reverse-causation). The use of MR in the context of DNA methylation and nsCL/P, will be discussed in more detail in **Chapter 4**.

An example of a non-conventional application of MR is when estimating the causal effect of liability to a disease on an outcome. The inheritance of dichotomous complex traits, such as disease status, can be modelled on the liability scale. Across

a population, liability to a disease can be assumed to be normally distributed; individuals over a liability threshold have the trait and individuals under do not [156]. The liability scale is easily demonstrated with psychiatric traits such as schizophrenia and autism, which are highly continuous traits but can be dichotomised into cases and non-cases. Psychiatric tests can be thought to be an estimate of disease liability; sub-clinical variation in psychiatric traits is present across the general population and threshold cut-offs are used for diagnosis of clinical cases. There is some evidence that sub-clinical liability to psychiatric traits may have beneficial effects, which could explain the evolutionary persistence of clinical psychiatric disorders. For example, genetic variants associated with autism spectrum disorder (ASD) have been shown to be associated with increased cognitive ability in the general population while genetic variants associated with increased educational attainment have been shown to be over-transmitted to children with ASD [157,158].

The liability model may be particularly relevant to nsCL/P because unaffected family members have been observed to have increased incidence of related sub-phenotypes [95]. Whether a liability model of inheritance is appropriate for nsCL/P will be discussed in more detail and explored in **Chapter 3**. In the context of demonstrating a causal relationship between liability to a trait and an outcome in an MR analysis, identified genetic variants associated with the trait are used as a genetic proxy for the genetic and non-genetic contributions contributing to liability. A consistent effect of genetic risk variants for the exposure on the outcome would support the liability hypothesis. However, if the effect of genetic risk variants for the exposure on the outcome is heterogeneous, then this suggests that the association may be driven by shared risk factors, such as environmental factors or specific biological pathways, or suggest weak genetic instruments for the exposure. The use

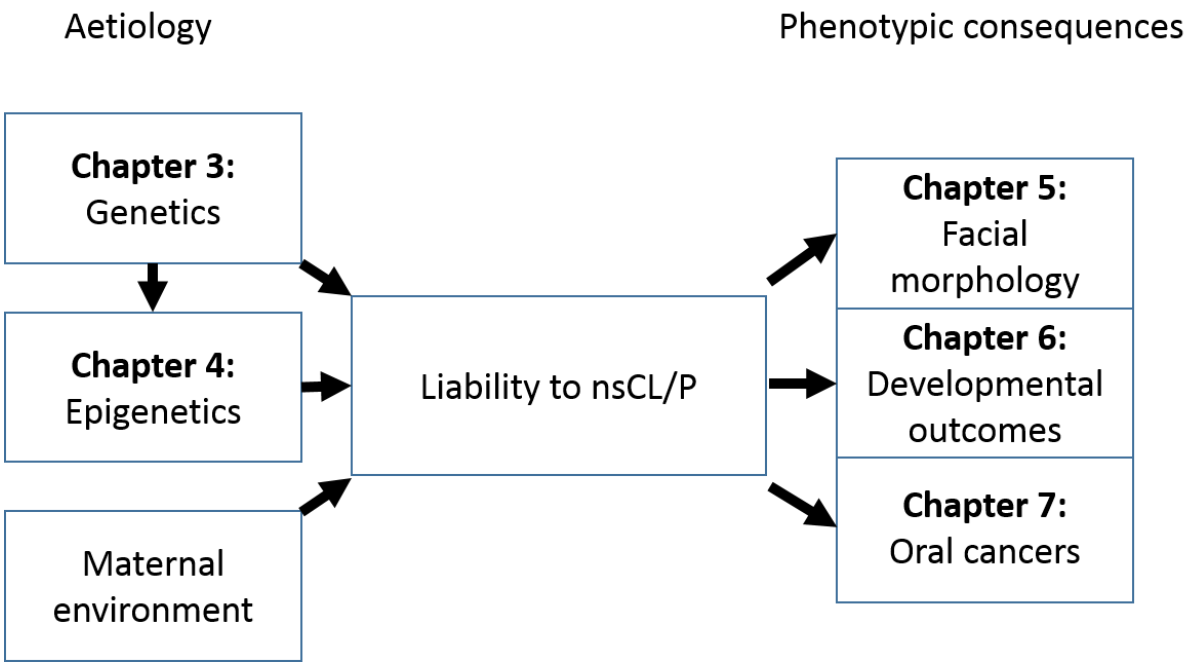
of MR to infer causal effects of liability to nsCL/P on phenotypes will be employed and discussed in **Chapters 5 and 7**.

1.3 Summary and overview of thesis aims

Under the right conditions, PRS and MR can be effective tools for answering epidemiological questions relevant to nsCL/P. However, both the use of PRS to detect genetic overlap, and the non-conventional use of MR to test the causal effect of liability to nsCL/P on phenotypes rely on certain assumptions about the dimensionality and genetic architecture of nsCL/P. Therefore, in **Chapter 3**, the aim is to explore the different components of the genetic architecture of nsCL/P and estimate the proportion of phenotypic variation explained by single nucleotide polymorphisms (SNPs), base-pair changes with a minor allele frequency (MAF) greater than 1%, on the genotyping chip.

In later chapters, the aim is to use PRS and MR to tease apart correlation and causality. In **Chapter 4**, epigenetic and genetic data are used in conjunction with MR to test the hypothesis that DNA methylation may mediate the effect of genetic risk variants for nsCL/P. In **Chapter 5**, PRS and MR are used to disentangle the shared genetics of nsCL/P and normal-range variation in facial morphology. In **Chapter 6**, PRS are used to explore if nsCL/P genetic risk variants explain the increased incidence of developmental outcomes amongst children with an OFC. Finally, in **Chapter 7**, PRS and MR are used to investigate shared genetic aetiology between nsCL/P and oral cavity/oropharyngeal cancer (**Figure 3**).

Figure 3: Flowchart of thesis analyses



Chapter 2: Data sources

2.1 Introduction

The work in this doctoral thesis utilises a range of different data types, obtained from many different studies and populations (**Table 1**). Many data sources are used in multiple chapters, so the purpose of this chapter is to give an overview of the different studies used in analyses and discuss primary data sources in detail for reference when reading subsequent chapters. Methodology that is specific to particular chapters will be described in the relevant chapter.

Table 1: Summary of data sources

Study	**Summary**	**Relevant data utilised in this doctoral thesis**
International Cleft Consortium	Orofacial cleft parent-offspring trio study	Individual level genotype data Case-control phenotyping
Avon Longitudinal Study of Parents and Children	Cohort study	Individual level genotype data DNA methylation data Phenotype data
UK Biobank	Cohort study	Individual level genotype data Phenotype data
Oral cavity and oropharyngeal cancer case-control data-set	Cancer case-control GWAS study	Individual level genotype data Case-control phenotyping
Bonn-II study	Orofacial cleft case-control GWAS study	GWAS summary statistics
Genetics of Overweight Young Adults	Case-cohort study of extreme BMI	Individual level genotype data DNA methylation data
The Cleft Collective	Orofacial cleft family study	DNA methylation data Case-control phenotyping
methWAS cohort	Orofacial cleft case-control study of DNA methylation	Summary statistics
Genotype-Tissue Expression project	Study of genotypic effects on gene expression	Summary statistics
NESDA NTR Conditional eQTL catalog	Study of genotypic effects on gene expression	Summary statistics
3D Facial Norms Database	Genetics of facial morphology study in the general population	Summary statistics Individual level data
1000 Genomes reference panel	Trans-ancestry genetic sequencing study	Individual level data

2.2 International Cleft Consortium (ICC)

2.2.1 Background

The International Cleft Consortium (ICC) is a large, publicly available (via dbGAP study accession: phs000094.v1.p1 [159,160]) database of genetic data relevant to OFCs. The data-set consists predominantly of OFC cases and their parents, sampled across a wide array of geographical locations in North America, Europe and Asia. Phenotype data consists of OFC subtype classifications as well as information on measured common maternal exposures (available in a subset of the sample). In total, the available data-set with complete genotype data includes 2,029 parent-offspring trios, 401 parent-offspring pairs, 88 singletons and 25 assorted extended families. Of the 2,543 individuals with a diagnosed OFC; 1,988 were classified as nsCL/P cases, 582 were classified as CPO cases and 21 presented with an unknown cleft subtype. Individuals were recruited from across 13 different centres (**Table 2**). Data from the ICC data-set was used in some form, in all results chapters.

Table 2: Recruitment centres and self-reported ethnicity of the ICC sample

Recruitment centre	European ancestry	East Asian ancestry	Other ancestry[1]	**Total**
Chengdu	0	452	0	**452**
Denmark	148	0	0	**148**
Iowa	288	1	2	**291**
Korea	0	198	0	**198**
Maryland	451	6	52	**509**
Norway	1173	14	9	**1196**
Philippines	0	0	292	**292**
Pittsburgh	407	0	5	**412**
Singapore	27	332	4	**363**
Taiwan	0	916	0	**916**
Utah	736	6	36	**778**
Weifang	0	843	0	**843**
Wuhan	0	691	0	**691**
Total	**3230**	**3459**	**400**	**7089**
Percent	**45.6%**	**48.8%**	**5.6%**	**100%**

1 Other ancestry includes the following classifications: African Americans, American Indians and Pacific Islanders

2.2.2 Phenotyping

OFC cases were classified as having CLO, CL/P or CPO through either a treatment centre or population-based registry. A subset of parents were interviewed regarding family history and maternal exposures during the peri-conceptual period (3 months prior to conception through the first trimester) and information was derived on the offspring's exposure to maternal smoking, alcohol consumption and multivitamin use during pregnancy [160].

2.2.3 Genotyping and quality control (QC)

ICC tissue samples were collected from whole blood (83.1%), buccal brush/swab (11.0%), saliva (2.8%), mouthwash (1.8%) and dried blood spots (1.2%). DNA samples were extracted using methods varying across recruitment centre and tissue source. DNA samples were genotyped using the Illumina Human610 Quadv1B array and the BeadStudio calling algorithm at the John Hopkins Center for Inherited Disease Research (CIDR).

Of 7,347 DNA samples from study subjects genotyped using the Illumina Human610 Quadv1B array SNP genotyping platform, scans from 7,089 subjects passed QC for unexpected relatedness, gender errors (where self-reported gender estimated from X chromosome heterozygosity rates is inconsistent with the genotype sample) and missingness (>5%). This sample was released on dbGAP. Pre-dbGAP release, SNPs in sample-chromosome combinations with a chromosomal anomaly (e.g. aneuploidy) were also excluded. Post dbGAP release, SNPs were excluded for missingness (>5%), MAF (<5%) and deviation from Hardy-Weinberg Equilibrium (HWE) ($P < 0.05$) using PLINK [161] leaving genotype data for 490,493 SNPs.

2.2.4 Strengths and limitations

Two major characteristics of the ICC data-set are the study design (i.e. parent-offspring trios) and the ancestral heterogeneity of study participants. The sampling of parent-offspring trios has advantages regarding the effect of population stratification bias and allows the testing of parent of origin effects. However, relevant to analyses, the recruitment of trios is largely a disadvantage because many planned analyses required an unrelated control group for comparison.

The ancestral heterogeneity of the ICC data-set is a major advantage in that it allows replication in different ancestral groups, which was utilised in **Chapter 3**. However, for the purposes of analyses the heterogeneity, even amongst self-reported Europeans, increased the difficulty of case-control matching in **Chapter 3**. Similarly, many analyses in this thesis required nsCL/P variants identified in European populations, so the non-European ICC samples were not included in many analyses.

Beyond the limitations regarding my analyses, the ICC is a rich resource for exploring the genetics of OFCs. The ICC is currently the largest publicly available collection of OFC cases and family members, including both phenotype and genotype data on over 2,500 OFC cases.

2.2.5 Analysis subsets

2.2.5.1 European nsCL/P parent-offspring subsets

A subset, of European nsCL/P cases and parental controls, was created for a pedigree analysis in **Chapter 3** Firstly, the subset was restricted to pedigrees consisting of an offspring with an OFC with at least one parent in the data-set. Secondly, the subset was restricted to pedigrees where the offspring were phenotyped as either CL/P or CLO. Finally, the subset was restricted to pedigrees of

self-reported European descent. The final sample consisted of 638 parent-offspring trios and 178 parent-offspring pairs. The meta-analysis GWAS summary statistics generated using this sample were used in all five subsequent results chapters.

For a specific analysis in **Chapter 3** requiring parent-offspring data, parents with a diagnosed OFC were also removed from analysis. The final sample consisted of 604 parent-offspring trios and 198 parent-offspring pairs.

2.2.5.2 Asian nsCL/P parent-offspring subset

For a specific analysis in **Chapter 3**, parent-offspring samples of East Asian descent were required. As for the European subsets, a sample of Asian nsCL/P cases and parental controls was created by restricting the data-set to pedigrees of self-reported East Asian descent where the offspring were phenotyped as CL/P or CLO and removing parents with a diagnosed orofacial cleft. The final sample consisted of 759 parent-offspring trios and 159 parent-offspring pairs.

2.2.5.3 European nsCL/P cases only subset

A European nsCL/P case only sample was utilised in matched case-control analysis in **Chapter 3**. The subset was restricted to individuals of self-reported European descent with a CL/P or CLO phenotype. The final sample consisted of 838 nsCL/P cases.

2.3 Avon Longitudinal Study of Parents and Children

2.3.1 Background

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a longitudinal birth cohort study based in the former county of Avon in the United Kingdom that recruited pregnant women with expected delivery dates between 1 April 1991 and 31 December 1992. The initial number of enrolled pregnancies was

14,541, which resulted in 14,062 live births and 13,988 children alive at the age of 1. When the oldest children in the study were around 7 years of age, the initial sample was boosted with eligible cases who had failed to join the study originally. Data from ALSPAC participants were used in **Chapter 5** and **Chapter 6**.

2.3.2 Phenotyping

ALSPAC is a deeply phenotyped cohort, with self-report questionnaires and/or clinic sessions used to collect phenotype data from the mother and her partner (both during pregnancy and post birth) and the children (post birth), at many different time-points. The phenotypes of relevance for this thesis are outcomes potentially relevant to nsCL/P. Facial phenotypes derived from 3D facial scans of ALSPAC participants in a clinic session at age 15 were utilised in **Chapter 5**. Hearing assessment phenotypes (audiometry and tympanometry), speech assessment, anthropometric measures and dental outcomes were utilised in **Chapter 6**.

2.3.3 Genotyping and QC

9,912 ALSPAC children were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform. Individuals with gender errors, excessive or minimal autosomal heterozygosity (where the proportion of genotyped markers with different alleles is higher or lower than expected, which may suggest possible sample contamination), disproportionate levels of individual missingness (>3%), unexpected relatedness (>10%) or evidence of non-European ancestry (which was determined by comparing study participants to individuals of European ancestry from the HapMap 2 reference panel) were excluded from analysis.

The samples were pre-phased, which involves statistically estimating individual's haplotypes, using ShapeIT (v2.r644) [162] a programme that utilises relatedness during phasing. The sample was then imputed, where missing

genotypes are estimated using a reference panel, to the 1000 Genomes (Phase 1, Version3) [163], which is described in **Chapter 2.6.8**, using IMPUTE3 V2.2.2 against the reference panel. After QC for SNPs on imputation quality (INFO< 0.8) and MAF, 8,099,747 SNPs and 8,860 individuals were available for analysis. ALSPAC genotype data were used in **Chapter 5** and **Chapter 6** to construct nsCL/P PRS.

2.3.4 DNA methylation

As part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) project [164], genome-wide DNA methylation data were generated for 1,018 ALSPAC mother-child pairs at five different time-points. Bisulfite sequencing, which converts non-methylated cytosine nucleotides to uracil allowing detection of methylated cytosines, was used in conjunction with the Illumina Infinium HumanMethylation450K BeadChip assay to quantify DNA methylation at over 480,000 CpG (where a cytosine nucleotide is followed by a guanine nucleotide) sites across the genome. After QC and normalisation [165], data were reported as methylation beta values, ranging from 0 (completely unmethylated) to 1 (completely methylated). Analyses in **Chapter 4** used the ARIES methylation data generated from the offspring cord blood samples, which were collected at birth.

2.3.5 Strengths and limitations

The main advantage of the ALSPAC cohort is the availability of detailed data on a wide-range of phenotypes, along with both genetic and DNA methylation data. The availability of hearing, speech and facial morphology phenotypes (all derived by specialists) is relatively unique and allows for the testing of genetic overlap between these phenotypes and nsCL/P. Arguably, a disadvantage of ALSPAC, is the modest sample size compared to large Biobanks, because genetic analyses can often require large sample sizes to be well-powered.

2.4 UK Biobank

2.4.1 Background

The UK Biobank is a large-scale cohort study of 502,655 participants aged between 40-69 years. Study participants were recruited from 22 recruitment centres across the United Kingdom between 2006 and 2010 [166]. Data from UK Biobank participants were used in **Chapter 3** and **Chapter 7**.

2.4.2 Phenotyping

Questionnaires completed at baseline were used to extract information on a wide-range of phenotypes for the majority of study participants. Phenotypes relevant to my analyses were; ICD10 codes from hospitalisation events, and self-reported alcohol consumption/tobacco smoking phenotypes which were used in analyses in **Chapter 7**.

2.4.3 Genotyping and QC

488,377 UK Biobank participants were assayed using two very similar genotyping arrays, the UK BiLEVE Axiom™ Array by Affymetrix1 (N= 49950) and the closely-related UK Biobank Axiom™ Array (N= 438427). Directly genotyped variants were pre-phased using SHAPEIT3 [167] and then imputed using Impute4 using the UK10K [168], Haplotype Reference Consortium [169] and 1000 Genomes Phase 3 [163] reference panels. Post-imputation, data were available on approximately ~96 million genetic variants [166,170]. For the purposes of analyses undertaken in **Chapter 3**, 152,249 individuals from the first genotype data release in May 2015 were used. For analyses in **Chapter 7**, the full sample from the second genotype data release in July 2017 (N=488,377) was used.

2.4.4 Strengths and limitations

The main strength of the UK Biobank is the size of the data-set; genotype data and phenotype data on almost 500,000 individuals is unprecedented, with obvious advantages for the statistical power of analyses. UK Biobank study participants are mostly of western European descent, which limits the possibility of matching Biobank controls to cases of non-European descent, but the large sample size allows the potential for accurate case-control matching for cases of recent European ancestry in **Chapter 3**. Two limitations of the UK Biobank relevant to my analyses are the absence of phenotypes relating to nsCL/P related outcomes, and the potential for selection bias. The UK Biobank study participants are middle-aged, so phenotypes related to outcomes in children such as speech and hearing are not available, while selection into the study has been proposed as a potential source of bias [171].

2.5 Oral cavity and oropharyngeal cancer (OC/OPC) data-set

2.5.1 Background

The OC/OPC data-set refers to an amalgamation of OC/OPC cases and controls from different studies, used in a previous GWAS [172]. The data-set includes 6,034 cases and 6,585 controls from 12 epidemiological studies, mostly hospital-based case-control samples, from North America, South America and Europe. The majority of participants were from the International Head and Neck Cancer Epidemiology Consortium (INHANCE), the European Prospective Investigation into Cancer and Nutrition (EPIC) or the Head and Neck 5000 (HN5000). OC/OPC data were used in **Chapter 7**, where potential genetic overlap between nsCL/P and OC/OPC was investigated.

2.5.2 Phenotyping

OC/OPC is a heterogeneous phenotype, including cases with cancer of the oral cavity, oropharynx, hypopharynx, or at multiple sites. The OC/OPC subtypes (oral cavity, oropharyngeal, hypopharyngeal and overlapping at multiple sites) were identified using ICD codes from hospitalisation events [172].

2.5.3 Genotyping and QC

Genotyping and QC has been described in detail previously [172]. In brief, DNA from blood or buccal cells was genotyped using the Illumina OncoArray, a custom cancer array. The majority of the samples were genotyped using a specific oral and pharynx cancer OncoArray but some of the shared controls were genotyped using a Lung OncoArray. Genotype calls were made using GenomeStudio software and a standardized cluster file for OncoArray studies. PLINK [161] was first used to exclude samples and SNPs with excessively high missingness (>20%). After the initial exclusions, in a second round of QC, samples and SNPs with missingness (>5%) were removed. Next, samples with chromosomal errors, excessive or minimal autosomal heterozygosity, unexpected relatedness (estimated relatedness > 0.3) and expected experimental duplicate pairs were also removed (the removal of controls was prioritised over cases). After QC, data on 513,311 genetic variants remained.

The study population was highly heterogeneous, with samples from Europe, North American and South America. Therefore, the data-set was divided into the three geographic regions and SNPs deviating from HWE ($P<1x10^{-7}$) were removed. Principal-components analysis (PCA), a statistical method of identifying trends in data which is discussed in more detail in **Chapter 3**, was used on a set of independent common markers in EIGENSTRAT [173] to identify 139 population

outliers, which were removed. STRUCTURE 2.3.4 [174] was used in conjunction with samples from the HapMap reference panel to determine the relevant ancestry (e.g. the code CEU refers to individuals of northern European ancestry sampled from Utah in the USA) for each individual. Finally, the directly genotyped data were imputed using the Michigan Imputation Server. SHAPEIT [175] was used for pre-phasing, with Minimac3 [176] used for imputation and the Haplotype Reference Consortium panel [169] used as a reference panel.

2.6 Secondary data sources

2.6.1 Bonn-II study

The Bonn-II study [23] was a central European based, case-control GWAS of nsCL/P, which including 401 nsCL/P cases and 1,323 controls genotyped using Illumina BeadChips (the Human610-Quad and the HumanHap 550k). GWAS summary statistics on 496,240 SNPs from the post quality-control, discovery sample of 399 nsCL/P cases and 1,318 controls were kindly made available by the principal investigators of the study. In **Chapter 3**, these summary statistics were meta-analysed with data from the ICC to attempt to replicate the summary statistics from a previous meta-analysis GWAS [24]. The summary statistics were used in subsequent analyses in all results chapters. The Bonn-II data were also used in **Chapter 3** to explore genetic architecture using PRS.

2.6.2 Genetics of Overweight Young Adults

The Genetics of Overweight Young Adults cohort (GOYA) [177] used a case-cohort sampling design to sample cases with extreme BMI scores and compared them to controls randomly sampled from the same cohort. Participants were sampled from the Danish National Birth cohort and the draft board examination cohort for men. Genotype and cord blood DNA methylation data were available for 1,000

children. Epigenetic and genotype data as well as estimated cell counts estimated using the Houseman method [178,179], ancestry principal components and DNA batch were available for 889 children. GOYA data are used in **Chapter 4** as a replication cohort for analyses in in ALSPAC.

2.6.3 The Cleft Collective birth cohort study

The Cleft Collective birth cohort study (CC) [180] recruited children born with an OFC in the UK between 2013 and 2016. Family members were invited to take part and data were collected on demographics and lifestyle via questionnaires. Blood and non-discarded lip and palate samples were collected during surgery. Surgical forms were used to phenotype OFC cases as CPO, CLP or CLO.

A subsample of 150 OFC cases (with no other known anomalies) were randomly selected and stratified by subtype: 50 CLP, 50 CPO and 50 CLO. Children have not been classified as syndromic because they have not been diagnosed as having any other anomaly, although because of the young age of the children, the non-syndromic status cannot be confirmed. The orofacial tissue type was dependent on the OFC subtype; lip samples were available for children with CLO and palate samples were available for children with CPO. Of the 50 children with CLP, 43 contributed a lip sample and 7 contributed a palate sample. Genome-wide DNA methylation was measured using the Illumina Infinium HumanMethylation450 BeadChip platform and functional normalisation was performed on the blood and tissue samples together. Of the original 300 samples, three blood and two lip samples failed QC. Surrogate variables were generated using the sva package in R to capture variation in the methylation data associated with technical batch and cellular heterogeneity [181 8]. Methylation data from the CC was used in **Chapter 4** to compare specific probe methylation in blood with methylation in lip and palate tissue.

2.6.4 methWAS cohort

The methWAS cohort included samples from 67 individuals diagnosed with nsCL/P and 59 age and sex matched controls, all of Brazilian ancestry. The average age at sampling was 5.29 years for cases and 6.45 years for controls. Whole-blood DNA was extracted from by the North Thames Regional Genetics Service with a subset of the sample (N=18) having available lip tissue samples recovered from surgery. DNA samples from the methWAS cohort were subjected to bisulfite conversion using the EpitectBisulfite Kit (QIAGen) and genome-wide DNA methylation was measured using the Illumina HumanMethylation 450 K Bead-Array platform. An epigenome-wide association study (EWAS), which estimates the association between DNA methylation and a phenotype of interest, was performed using the methWAS methylation data, with 11 CpG sites targeted for replicated in an independent UK sample of 171 cases and 177 controls [36]. Data from the methWAS cohort was used in **Chapter 4** to compare probes of interest from primary analysis with the results in the EWAS.

2.6.5 The Genotype-Tissue Expression project

The Genotype-Tissue Expression (GTEx www.gtexportal.org)) project is a resource database and tissue bank developed to aid the understanding of the relationship between genetic variation and gene expression in humans. The database includes information on expression quantitative trait loci (eQTL), which are genetic variants associated with gene expression, generated using genotype and RNA sequencing gene expression data from 43 distinct tissue types from 175 individuals [182,183]. In **Chapters 4** and **5**, SNPs of interest were looked up in GTEx to explore potential biological mechanisms related to gene-expression.

2.6.6 NESDA NTR Conditional eQTL Catalog

The NESDA NTR Conditional eQTL Catalog (NNC) is a repository of eQTL in whole blood, generated using genotype and gene expression microarray data from 4,896 individuals across two Dutch biobanks. Conditional eQTL analysis, which involves accounting for the correlation between nearby SNPs to estimate likely causal variants, was applied to distinguish between dependent and independent eQTL [184]. The catalogue is available at (https://eqtl.onderzoek.io/index.php?page=info). In **Chapter 4**, SNPs of interest were looked up in the NNC to explore potential biological mechanisms related to gene-expression.

2.6.7 3D Facial Norms Database

The 3D Facial Norms Database (3DFN) is a database of controls for craniofacial research with genetic data on 2,447 individuals, aged between 3 and 40 years of recent European descent. 2,272 individuals were recruited from Pittsburgh, Seattle, Houston or Iowa City as part of the 3DFN and the remaining 175 individuals were recruited as healthy controls for a separate study at Pittsburgh on orofacial cleft genetics. Study participants were screened for a history of craniofacial conditions and 3D-derived anthropometric measurements, 3D facial surface images and genotype data were derived from each participant [95,185]. Notably a GWAS of normal-range variation in facial morphology was published using the 3DFN cohort [95]. 3DFN data were used in **Chapter 5** as a replication cohort for facial morphology related analysis in ALSPAC.

2.6.8 Reference panels: The 1000 Genomes reference panel

Several different reference panels (HapMap2, 1000 Genomes, Haplotype reference consortium) have been mentioned earlier in the chapter when describing

genotype data and QC. Reference panels typically consist of a number of deeply sequenced individuals from different ancestral populations. There are several important uses of reference panels; first, it is useful to compare genotyped individuals to individuals from a reference panel to infer ancestry, second, reference panels can be used in genomic imputation to estimate markers that were not directly genotyped, and third, reference panels can be informative about linkage disequilibrium (LD), a characteristic of the genome where nearby markers are often correlated on a population level, which can be informative for identifying a proxy SNP for an unavailable SNP.

In several subsequent chapters, I use the 1000 Genomes [163] (which published results in 2015) as a reference panel. The 1000 Genomes data-set includes 2,504 deeply sequenced individuals from 26 different world-wide populations with genotypic data on over 88 million genetic variants [163]. Applications in later chapters, used the CEU individuals and the GBR individuals (Individuals of European descent from Great Britain) to estimate the MAF of specific variants in a European population and estimate LD to identify proxy SNPs and generate independent sets of SNPs.

Chapter 3: Exploring the genetic architecture of nsCL/P

3.1 Abstract

Understanding of the heritability and genetic architecture of nsCL/P is an important prerequisite for the interpretation of analyses in future chapters. Therefore, in this chapter I use available data-sets in conjunction with a variety of distinct statistical methods to estimate the SNP heritability of nsCL/P and make inferences about the genetic architecture.

Triangulating the results from different methods, strong evidence was found that nsCL/P is a highly polygenic trait, with common genetic variation on genotyping chips estimated to contribute between 20 and 33% of the phenotypic variance. The evidence for polygenic architecture suggests that PRS may be effective genetic proxies for liability to nsCL/P and supports phenotypic dimensionality of nsCL/P. Beyond implications for the genetic architecture of nsCL/P, analyses in this chapter have implications for the utility of case-control matching and for the use of different SNP heritability estimation methods. Firstly, ancestral matching of cases to controls from different studies was shown to be a non-trivial undertaking and should be considered carefully. Secondly, results and simulations suggested that the different SNP heritability estimation methods vary in effectiveness for samples affected by batch.

3.2 Introduction

In later chapters, I use genetic proxies for liability to nsCL/P in MR and PRS analyses in order to investigate possible causes and effects of nsCL/P. At the time of analysis (prior to a recent publication of a SNP heritability estimate for nsCL/P [26]),

there was limited knowledge about the genetic architecture of nsCL/P and no published SNP heritability estimates. As will be discussed in this chapter, the genetic architecture and SNP heritability of nsCL/P have important implications for analyses in later chapters and more generally, for the biological understanding of the trait aetiology.

The primary aims of this chapter were to estimate the SNP heritability and explore the genetic architecture of nsCL/P; for example, evaluating the hypothesis that nsCL/P has a highly polygenic architecture. In the introduction to this chapter, I describe the different components of genetic architecture and the relevant methods for investigating these components. Possible implications for analyses in later chapters as well as previous research on nsCL/P are also discussed.

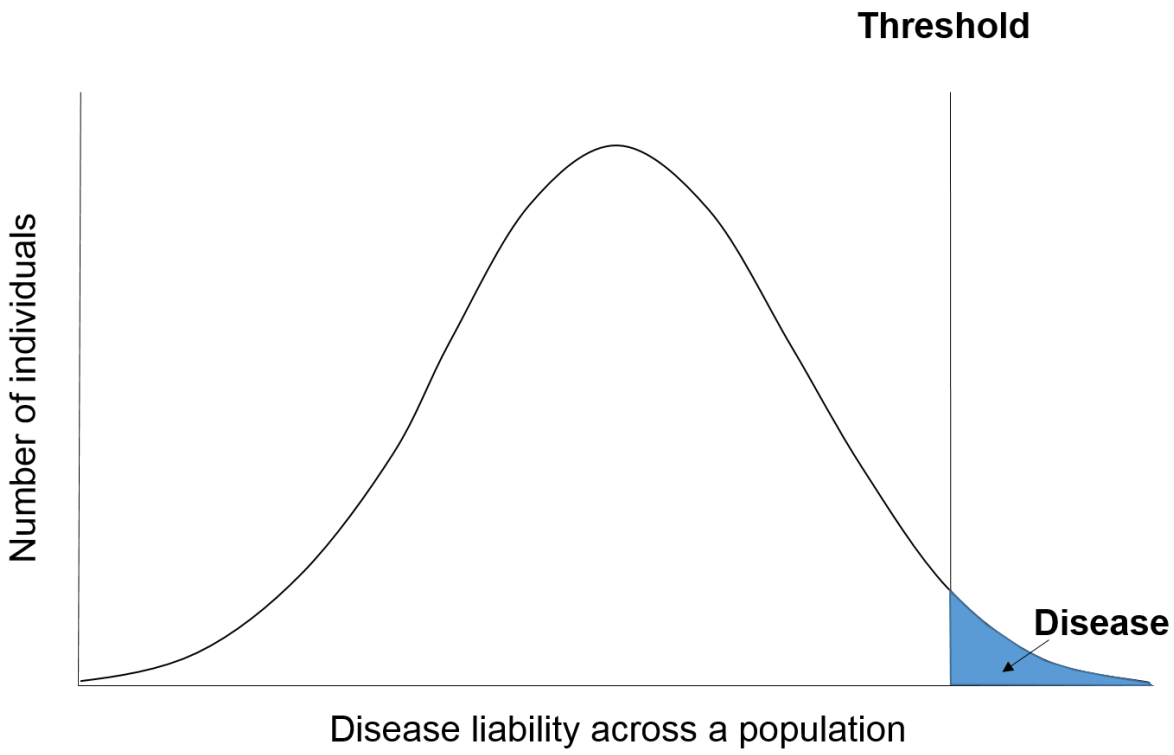
3.2.1 Heritability

The aetiological relevance of genetic factors to phenotypes can vary greatly across traits; height is highly genetic while an individual's first language is largely non-genetic (for example, consider the scenario where two identical twins were separated at birth and one grew up in France and the other grew up in Germany). Heritability is defined as the proportion of phenotypic variation attributable to genetic variation. The broad-sense heritability, denoted H^2, is the ratio of total genetic variance to total phenotypic variance $\frac{V_G}{V_P}$, the narrow-sense heritability, denoted h^2 considers only additive genetic variance $\frac{V_A}{V_P}$ while SNP heritability estimates include only additive genetic contributions from SNPs that are present on the assaying platform[186].

Heritability can be measured on the observed or liability scale. The observed scale is, as defined above, the proportion of phenotypic variation explained by

genetic factors across a population. Contrastingly, on the liability scale for heritability, phenotypic variation or disease liability is modelled as being continuous and normally distributed on a population level; individuals over the liability threshold have the trait and individuals under do not, with the proportion of the normal distribution over the threshold equal to the trait prevalence [186,187] (**Figure 4**).

Figure 4: Liability scale of phenotypic variation



For binary traits, the liability scale is typically preferred. This is because the phenotypic variation of a binary trait is dependent on the trait prevalence, and case-control studies will often have a higher prevalence of cases than the general population. Converting estimates from the observed scale to the liability scale allows

for heritability estimates for diseases with different prevalences to be compared, assuming that the liability model holds [187].

3.2.2 Genetic architecture

If genetic factors have a substantial role in a trait's aetiology, it is important to consider the genetic architecture, which is defined as the characteristics of genetic variation contributing to the aetiology of a trait. These include; the number of relevant genetic variants, their effect sizes, their allele frequencies as well as interactions between genetic variants (epistasis) or between the genome and the environment [146]. Inferences about the components of genetic architecture for nsCL/P may have important implications for the PRS and MR analyses in later chapters. Here, I describe the different components of genetic architecture and their importance.

3.2.2.1 Number of genetic risk variants

Unlike Mendelian traits, which are determined by genetic variation at a single locus, a complex trait is characterised by having multiple distinct genetic risk loci. The number of risk loci varies between complex traits, but many complex traits are thought to be affected by hundreds or thousands of independent genetic variants. Indeed, there is evidence that association signals for many complex traits are spread across the majority of the genome, which prompted a proposed omnigenic model of complex traits where the majority of the genome is relevant to trait aetiology via interconnected gene regulatory networks [145].

There are several important implications of the number of genetic risk variants influencing a trait which are pertinent to work in later chapters. Firstly, PRS which are used in later chapters to detect genetic overlap, are most effective for traits where many genetic variants contribute to trait aetiology. Secondly, for MR analyses, a large number of genetic instruments increases the power of sensitivity analyses

testing for horizontal pleiotropy [147,148]. Thirdly, the number of identified variants may support or oppose theories about the inheritance pattern of nsCL/P, e.g. the multifactorial liability or major risk gene models.

3.2.2.2 Frequency of genetic risk variants

Genetic variants vary in frequency across a population dependent on several factors including genetic drift, selection and the age of the variant. The frequency of disease predisposing genetic variants may be related to the disease prevalence. The common disease, common variant hypothesis [188,189] suggests that for prevalent complex traits, disease-causing alleles of small effect will be found commonly across the general population and that there is often only one disease-causing allele at a particular locus [190]. A well characterised example, is the discovery of a single common allele in the *APOE* locus that increases risk of both Alzheimer's and heart disease [191,192]. Contrastingly, proponents of the common disease, rare variant hypothesis argue that substantial variation in common diseases is related to rare mutations [193].

The frequency of relevant genetic risk variants can have important implications. Firstly, the frequency of genetic risk variants for a trait can be an important consideration for study design; standard genotyping chips have good coverage of common variation but mapping rare risk loci may require sequencing. Secondly, allele frequency is important for risk prediction screening. The effect of a variant on a population level is largely dependent on the allele frequency across the population of interest; a common variant with a small effect may be more useful for population level risk prediction than a rarer variant with a larger effect [146].

3.2.2.3 Magnitude of effect of genetic risk variants

The effect of a genetic variant is determined by the variants effect on the DNA sequence. For example, a missense mutation refers to a base-pair change within a protein-coding region which alters the amino acid in a protein. A missense mutation in *ADH1B,* which is involved in the production of an enzyme involved in alcohol metabolism, has an effect on an individual's alcohol behaviour. Genetic risk variants for the same trait can vary greatly in effect size; each risk allele of rs9939609 in *FTO* is associated with a 0.36 kg/m^2 in BMI compared to an increase of 0.023 kg/m^2 for each risk allele of rs657452 in *AGBL4* [194,195].

Natural selection has important implications for the relationship between allele frequency and the effect of the variant. Selection pressure is dependent on the effect of the variant on fitness to sexually reproduce and the allele frequency, with the cumulative effect dependent on the age of the variant. Therefore, the cumulative effect of selection pressure is higher on common genetic variants, which are older, meaning that common genetic variants with a large effect on fitness are uncommon. Contrastingly, rare variants (which can be newer mutations), are under less selection pressure and so may be more likely to have large effects on fitness. However, whether this affects the allele frequency of risk loci for a phenotype depends on how the trait affects fitness to sexually reproduce.

The effect size of genetic variants is important for several reasons. Firstly, the power of MR and PRS analyses depends on the proportion of phenotypic variation explained by relevant genetic variants. Secondly, for screening, variants of large effect are (dependent on allele frequency) desirable for risk prediction. Thirdly, from an evolutionary perspective, variants with large effects can have important implications for human history. For example, a variant in *CREBRF*, that is extremely

rare in most populations, is strongly associated with increased BMI in Samoans. Each risk allele is associated with a 1.36-1.45 kg/m^2 increase in BMI, an effect size approximately four times that of the previously described variant in *FTO* [196]. This variant is thought to have been under positive selection in the past when food was scarcer but may now be under negative selection with rising obesity rates in the Samoan population.

3.2.3 How to estimate the heritability and characterise the genetic architecture of nsCL/P

The next step is to consider how one can estimate the heritability of nsCL/P and explore the different components of genetic architecture. There are several different study designs and methodologies that differ in terms of data requirements (e.g. genotype data or pedigrees without genotype data), output (e.g. some estimate heritability and some identify risk loci) and interpretation.

Before I describe the relevant methodologies, it is pertinent to begin by introducing the concepts of identical by descent (IBD) and identical by state (IBS). Assuming two individuals share an identical stretch of DNA, if the stretch is inherited from a recent common ancestor it is said to be IBD while conversely if the stretch cannot be established as being inherited from a recent common ancestor, it is said to be IBS. Individuals with recent common ancestors will share IBD stretches of chromosomes (haplotypes); closely related individuals will share long segments of DNA while less closely related individuals will share smaller segments.

3.2.4 Pedigree and twin studies

One way of investigating the contribution of genetics to the aetiology of a phenotype is to use phenotyped pedigrees. The premise is to compare the

phenotypic differences between related individuals in relation to the proportion of the genome shared between the relatives, i.e. the proportion of alleles that are IBD.

Twins are especially useful for exploring heritability because certain characteristics allow for a design akin to a natural experiment. Monozygotic (MZ) twins are genetically IBD, sharing ~100% of alleles, while dizygotic (DZ) twins share around 50% of alleles that are IBD, but both types of twins can be assumed to share a common environment. Twin studies contrast the phenotypic concordance of a trait in monozygotic and dizygotic twins. In theory, a heritable trait will be more concordant in monozygotic twins than in dizygotic twins, which in turn will be more concordant than between unrelated individuals. An estimate of the narrow-sense heritability can be approximated by a function of the difference between the MZ/DZ phenotypic concordance. Similarly, sibling comparisons can be used to estimate the narrow-sense heritability, by comparing the trait recurrence rates between siblings to the trait incidence in the general population.

Early twin studies demonstrated that nsCL/P is a highly heritable trait, with substantially higher concordance rates between MZ twins compared to DZ twins. In Brazilian and Scandinavian populations, twin heritability estimates ranged from 45 to over 90% [197-199]. However, there are several potential limitations of heritability estimates from twin studies. Firstly, the shared environment may not be the same between MZ and DZ twins (MZ twins may be treated differently to DZ twins) and secondly, twin concordance can be difficult to measure if ascertainment is incomplete. The interpretation of heritability estimates from twin studies is complex because it is impossible to estimate the contribution of additive, dominance or environmental differences and it is therefore unclear whether the twin heritability estimate is an accurate estimate of the narrow-sense heritability [186,187,200,201].

Prior to the availability of genotype data, there was considerable disagreement between pedigree studies about the most likely inheritance pattern of nsCL/P. Carter first proposed a multifactorial model of inheritance [156] by arguing that the recurrence rates of nsCL/P are inconsistent with both recessive and dominant models of Mendelian inheritance [202]. The existence of sub-phenotypes in unaffected family members and the results of some pedigree studies supported the multifactorial model [203-206], but the results of other pedigree studies suggested that the recurrence patterns were more consistent with autosomal major gene inheritance, potentially in conjunction with the environment [207-210]. The possibility of a major nsCL/P susceptibility gene prompted the use of linkage analysis to identify the responsible gene, which I will discuss in the next section.

3.2.5 Genotype-driven methods

Genotyping refers to the examination of an individual's DNA sequence; genotype data are commonly used to map genetic risk loci for traits by comparing the DNA sequence of different individuals but can also be used to estimate the heritability of genetic variants contained on the genotyping platform. The results of genetic mapping studies (e.g. the number of independent risk variants identified, the allele frequencies and the effect sizes) can be used to make inferences about the components of genetic architecture. A major caveat is that the different genetic mapping study designs have different strengths and weaknesses (e.g. linkage analysis within pedigrees is most effective for traits with a major susceptibility locus). Therefore, it is important to consider the results of the different study designs to make accurate inferences about the genetic architecture of nsCL/P.

3.2.5.1 Linkage analysis

Linkage analysis is a genetic mapping method that exploits the IBD sharing within pedigrees to measure the co-segregation of segments of DNA with a trait. During meiosis, homologous pairs of chromosomes line-up and undergo crossing over, where the maternal and paternal chromosomes recombine so that the resulting chromosomes contain a mixture of maternal and paternal genetic variation. It is well characterised that recombination between two genetic loci is inversely related to the distance between the loci on the chromosome and so loci that are close together are often inherited together [14].

Linkage was employed with some success to identify regions relevant to syndromic forms of cleft, e.g. Van der Woude syndrome and popliteal pterygium syndrome [15,16] but mapping risk loci for nsCL/P proved to be difficult. Multiple regions were found across different studies to co-segregate with nsCL/P, including loci on chromosome 6p23-24 (*F13A*) [17,18,211], chromosome 2p13-p14 and chromosome 19q13.1 (*BCL3*) [19] but findings were not consistently replicated [20]. Although effective at detecting rare alleles with high penetrance, linkage analysis has been shown to be largely ineffective at identifying higher frequency genetic variation with modest effects [14,212]. The difficulties mapping loci for nsCL/P using linkage were inconsistent with the existence of a major nsCL/P susceptibility gene.

3.2.5.2 Genome-wide association studies and Transmission Disequilibrium Test

An alternative method for mapping genetic risk loci for complex traits is association testing, such as in a GWAS. Association testing involves quantifying the statistical association between genotyped genetic variants and a phenotype, requiring only that alleles are IBS [213]. There are three possible reasons for an observed association between a genetic variant and a phenotype. The best-case

scenario is that the allele is a causal variant which directly affects disease susceptibility. The next best scenario is that, because some subjects share a recent common ancestor, the allele is in LD with a causal variant. The worst-case scenario is that the association is a false positive, potentially because of bias from population stratification or cryptic relatedness [214].

As alleles are not required to be IBD in association testing, the majority of GWAS compare cases to unrelated controls. However, cases with related controls (i.e. parents or siblings) can also be used for genetic association testing. The Transmission Disequilibrium Test (TDT) [215,216] measures the over-transmission of heterogeneous alleles from parents to their unaffected offspring and so can be described as detecting linkage in the presence of association. Notable advantages of sampling trios are that parental environmental risk factors can be phenotyped, the TDT is robust to population stratification and that parent-of-origin effects can be investigated. The main disadvantage of trio designs is lower statistical power; a parent-offspring trio has roughly equivalent power to a single case and a single control from a case-unrelated control design [214,217].

Both trios [21,30,71] and case-control [22,23] designs have been used for GWAS of nsCL/P. A primary reason for sampling trios is the interest in identifying maternal risk factors and gene-environment interactions relevant to nsCL/P. GWAS have been very successful in identifying risk loci for nsCL/P, with over 40 distinct genetic risk loci identified in studies across ethnically heterogeneous populations [9,21-24,27-30,218]. The existence of many common risk variants for nsCL/P supports a multifactorial model of inheritance for nsCL/P, but interestingly, many of the identified risk variants have substantial effect sizes. For example, the risk allele of rs987525 (a major nsCL/P risk SNP) has a MAF of around 25% but, uncharacteristically for a common

variant, is associated with substantially increased risk of nsCL/P (Relative risk > 2.0) [24]. Smaller but substantial effect sizes are observed for many other common nsCL/P risk loci, which may have implications for the selection pressure on nsCL/P related alleles.

A point to consider when using GWAS results to make inferences about genetic architecture is that the number of genetic variants identified often increases with sample size. This is because the main criteria in a GWAS is usually whether a variant passes the genome-wide significance threshold. As sample sizes tend to infinity, a huge number of variants will pass the genome-wide significant threshold for polygenic traits with smaller and smaller effects (dependent on allele frequency). For example, a recent GWAS of height in 700,000 Europeans reports 700 loci at genome-wide significance [219]. Therefore, the ability to make inferences about trait-specific genetic architecture can be dependent on available data-sets; the rarity of nsCL/P means that current GWAS are still modestly sized.

3.2.5.3 Using GWAS data to estimate SNP heritability and polygenicity

Genotype data can also be used to estimate the heritability of markers on the genotyping chip (often referred to as the SNP heritability). There are currently several different methods for estimating the SNP heritability, including; Genetic relationship matrix restricted maximum likelihood (GREML) as implemented in Genome-wide Complex Trait Analysis (GCTA) software [220], the use of Bayesian linear mixed-model approaches as implemented in BOLT-Linear mixed models (BOLT-LMM) [221], Linkage Disequilibrium Adjusted Kinship (LDAK) [222,223], Additive Variance Explained and Number of Genetic Effects Method of Estimation (AVENGEME) [224] and LD score regression [225], which utilises summary level GWAS data. Individual level genotype data from pedigrees can also be used to test the

polygenicity of a trait; the polygenic transmission disequilibrium test (PTDT) measures over-transmission of polygenic risk between unaffected parents and affected offspring [158].

The SNP heritability of nsCL/P has been previously estimated to be around 30% in a European population using GCTA, with 25% of the variance explained when restricting to 24 known risk loci [26]. However, there is some evidence suggesting that the SNP heritability of nsCL/P may vary across different populations. In a Chinese population, 26 known risk loci were found to account for around 11% of the heritability of nsCLP which is substantially lower than the estimate from the European population [26,28], perhaps because of differences in LD structures between the different populations. It is worth noting that because some of the known risk loci were identified in the same study populations used for the heritability estimates, the variance explained by the known risk loci may be inflated by winner's curse. Regardless, the variance explained by SNPs in the two populations suggests that common variation explains a large proportion of the phenotypic variation for nsCL/P.

3.2.5.4 Missing heritability

Although GWAS have identified thousands of risk loci across complex traits, SNP heritability estimates are consistently lower than heritability estimates generated from twin studies [226]. Indeed, the SNP heritability estimate of 30% for nsCL/P is substantially lower than the twin study estimates of 45-90%. Different theories have been proposed for these differences, including; the role of LD and MAF [222]; the role of rare variation [227,228]; the effects of copy number variation [229]; and biased heritability estimates from pedigree studies [230].

3.2.5.5 Genomic imputation, whole exome and whole genome sequencing

The lack of coverage of rare variation on a typical genotyping chip means that in general, the majority of variants identified in GWAS are relatively common in frequency. This means that GWAS results may under-represent the importance of rare variation in trait-specific genetic architecture. However, the number of identified low frequency risk variants is increasing with the use of sequencing and genomic imputation. Sequencing involves complete (or near-complete) genotyping of the DNA sequence in regions of interest, which is important for identifying rare variation missed by typical genotyping chips. Targeted sequencing involves sequencing a specific region of interest, whole exome sequencing involves sequencing the protein coding segments of the genome and whole genome sequencing involves sequencing of the majority of the genome [146]. Large deeply-sequenced reference panels (described in **Chapter 2**) can be used in genomic imputation, where the genotypes of non-genotyped variants are estimated using LD and phase from a reference panel. The imputation of GWAS results allows further identification of lower-frequency genetic risk variants.

Imputation and sequencing has led to increased discovery of low frequency and rare genetic variants for nsCL/P. Genomic imputation of a nsCL/P data-set identified 4 additional risk loci [26], suggesting that some lower frequency risk markers were missed by the original genotyping arrays. Targeted sequencing has identified rare genetic variation in known nsCL/P candidate genes that may contribute to disease aetiology [31,231] and whole exome sequencing has identified further candidate genes for nsCL/P, containing rare missense and deleterious variants [232,233]. These findings suggest that rare genetic variation has an important aetiological role and may explain some of the missing heritability, but larger sample sizes are required to

better characterise the effects of rare variation on nsCL/P. One caveat is that many of the syndromic forms of CL/P are often phenotypically indistinguishable from nsCL/P, complicating the search for rare variation. Currently, rare variation is unlikely to be useful for PRS and MR analyses because variants identified in sequencing studies are unlikely to be commonly genotyped or imputed in other studies.

3.2.6 Caveats and analysis plan

The primary aim of this chapter was to use a variety of different methods to make inferences about the genetic architecture and heritability of nsCL/P. However, there are two major caveats that require further background.

3.2.6.1 Construction of nsCL/P summary statistics

Firstly, it was necessary to independently generate nsCL/P GWAS summary statistics, used in this chapter and later chapters, because the summary statistics from the largest previously published meta-analysis GWAS of nsCL/P in Europeans [24] were not publicly available. This GWAS [24] was a meta-analysis of two previous studies, the ICC trios and the Bonn-II study [21,23]. Although the meta-analysis summary statistics were unavailable, I had access to the individual level genotype data from the ICC and GWAS summary statistics from the Bonn-II study and therefore was able to use the data from these studies to generate meta-analysis nsCL/P summary statistics.

Meta-analysing the results from the two studies requires careful consideration; the TDT design tests for linkage in the presence of association and commonly reports the proportion of transmitted risk alleles, while a case-control GWAS design tests for association and often reports an odds ratio (OR). However, OR of the associations between genetic markers and the phenotype in a TDT can be estimated

as a function of the proportion of transmitted high-risk alleles [234], allowing for harmonisation of results between the two studies.

3.2.6.2 Case-control matching

The second caveat is that the ICC data-set consists of parent-offspring trios but many heritability estimation methods, e.g. GCTA, require samples consisting of cases and unrelated controls. To circumvent this issue, one possibility was to match nsCL/P cases with population controls, but this was likely to be a challenging task because the nsCL/P cases (from the ICC), even those of European descent, were sampled from different recruitment centres across Scandinavia and the US. This meant that matching on ancestry to a single homogeneous population study could lead to population stratification bias, where systematic population differences between cases and controls lead to differences in allele frequency. Differences in disease prevalence across study populations combined with population stratification can lead to genetic associations that are unrelated to genuine biological differences [235]. A proposed solution to the ancestral heterogeneity of the ICC data-set was to match nsCL/P cases with population controls from the UK Biobank, which sampled a large number of ancestrally heterogeneous individuals.

Further considerations relevant to case-control matching are firstly, the possibility of batch effects between cases and controls, and secondly, the most appropriate way to match cases to controls. Batch effects are systematic genotyping differences caused by differences in the genotyping process, such as different genotyping chips, and can be difficult to distinguish from genuine biological differences. Batch effects are a particular problem in this instance as firstly, the ICC study and the UK Biobank used different genotyping chips, and secondly, all the

nsCL/P cases are on one genotyping chip and all of the proposed controls are on another.

In terms of matching cases to controls, attempting to pair cases with ancestrally homogeneous controls is an obvious starting point. One method that can be used to explore ancestry is PCA. PCA is a commonly used adjustment to account for population stratification in genetic association studies, the principal being that genome-wide IBS sharing between individuals identifies recent common ancestry [173,236]. PCA has limitations, notably it has reduced effectiveness for highly admixtured populations with uneven sampling [237,238]. In this instance, derived principal components from a merged ICC-UK Biobank sample are not necessarily measures of ancestry as they may pick up any form of systematic variation across a data-set such as batch effects.

An alternative approach is to use previously derived markers of ancestry, ancestral informative markers (AIMs), to infer ancestry. Using AIMs to infer ancestry is less computationally intensive than PCA and may be more likely to pick up biogeographical substructure. AIMs have been previously identified for 9 different populations from the 1000 Genomes [163] and can be used to generate admixture components inferring the ancestry of genotyped individuals [239 240].

3.2.6.3 Overview of analysis plan

First, meta-analysis GWAS summary statistics were constructed using available data sources. Second, case-control matching was used to pair nsCL/P cases from the ICC with UK Biobank controls. Third, relevant genetic architecture and heritability methods were used to make inferences about nsCL/P. Relevant data-sets and analyses are contained in **Table 3** in the next section. In this chapter, I performed all described analyses.

3.3 Materials and methods

3.3.1 Study participants

In this chapter, I utilised several datasets; the ICC individual level trio genotype data, the Bonn-II study GWAS summary statistics and the UK Biobank controls. The three data-sets have been described previously in detail in **Chapter 2**.

In brief, the ICC data-set consists of over 2,500 individuals with an OFC predominantly of European or Asian descent, with parental controls. The Bonn-II study was an nsCL/P case-control GWAS including 399 cases and 1,318 controls. The UK Biobank is a cohort study including 502,655 participants aged between 40-69 years, sampled from across the UK. More information on analyses and relevant data-sets is contained in **Table 3**.

Table 3: Different analyses and relevant data-sets as described in **Chapter 2**

Analysis	**Datasets used**
nsCL/P meta-analysis summary statistics	European ICC trios and Bonn-II summary statistics
Matched nsCL/P-control sample	European ICC nsCL/P cases and UK Biobank controls
Polygenic transmission disequilibrium test	European and Asian ICC trios with affected parents removed
Sibling recurrence rate	N/A (used estimates from literature)
LD score regression	nsCL/P meta-analysis summary statistics
Genome-wide Complex Trait Analysis	Matched nsCL/P-control sample
AVENGEME	Bonn-II summary statistics Matched nsCL/P-control sample

3.3.2 Generation of nsCL/P meta-analysis summary statistics

First, genetic loci identified with nsCL/P were identified using a TDT. PLINK [161] was used to run a TDT on a subset of the ICC data (**Chapter 2.2.5.1**), consisting

of 638 parent-offspring trios and 178 parent-offspring pairs of European descent. TDT effect sizes were reported as OR.

Second, the TDT summary statistics were meta-analysed with the Bonn-II summary statistics. In this instance, with only two studies to meta-analyse, and similar phenotyping and population structure in both datasets, a fixed effects model was assumed. METAL [241] was used to meta-analyse the OR from the TDT with the Bonn-II case-control GWAS [23] of 399 cases and 1,318 controls (**Chapter 2**) using a fixed-effects model. The final sample consisted of 1215 cases and 2772 controls, although it is worth noting that one parental control is not statistically equivalent to one unrelated control.

As the meta-analysis nsCL/P GWAS summary statistics and their components have been described in previous publications [21,23,24], biological implications of the GWAS results are omitted. The results are instead validated by comparing the P-values of the constructed GWAS summary statistics with the top hits from the published meta-analysis GWAS [24].

3.3.3 Case-control matching

3.3.2.1 Merging European nsCL/P cases and UK Biobank

Firstly, the 838 European ICC nsCL/P cases (**Chapter 2.2.5.3**) were merged with controls from the first genotype release of the UK Biobank (see **Chapter 2.3**). To increase computational speed when generating admixture components, the UK Biobank sample of over 150,000 individuals was split into 50 subsets at random. Each of the subset samples was then restricted to include only individuals of European ancestry, confirmed by PCA. The European nsCL/P and UK Biobank subsets samples were then filtered using a list of 130,025 AIMs from the GenoChip [239], of which 96,900 were genotyped in all samples. The nsCL/P cases were then

merged with each of the 50 UK Biobank subsets using PLINK. SNPs with mismatched alleles between the two data-sets or missingness greater than 2% across the combined sample were removed, leaving 69,219 AIMs to be used for the generation of admixture components.

3.3.2.2 Generating admixture components

Secondly, admixture components were derived in each of the merged nsCL/P-UK Biobank subsets. Admixture components were preferred over principal components because of the computational intensity of generating principal components and the possibility of picking up dimensions independent from ancestry in the combined sample. Nine admixture components, representing nine of the 1000 Genomes [163] worldwide populations, were identified in each of the 50 combined samples (the nsCL/P cases and a subset of the UK Biobank controls) by applying ADMIXTURE v1.3 [174] to the 69,219 AIMs. Admixture was applied in supervised mode against the nine previously curated gene pools [240].

3.3.2.3 Propensity score matching

Thirdly, each nsCL/P case was matched with 4 appropriate controls using propensity score matching which involves matching controls to cases, conditional on a defined set of covariates with an optimal matching defined as the matching with the minimum Euclidean distance [242]. Propensity score matching [242] in R, using the nearest neighbour algorithm, was used to match the nsCL/P cases to UK Biobank controls with the nine admixture components as covariates. Analysis was performed separately for each of the 50 subsets and the best 4 matched controls, in terms of minimum Euclidean distance, for each case were then taken forward to the analysis stage (matched controls were sampled without replacement).

3.3.2.4 Evaluating matching quality

Finally, the quality of the matching in the nsCL/P cases/matched UK Biobank controls sample was evaluated using principal components. To generate the principal components, the European nsCL/P and UK Biobank samples were first filtered to include only HapMap3 SNPs [243]. The files were then merged using PLINK, SNPs with mismatched alleles between the two data-sets were removed and the data-set was restricted to nsCL/P cases and matched UK Biobank controls. Strict QC was then performed, removing SNPs with MAF < 0.01, HWE P-value <0.01 and missingness >2%. Pairwise LD pruning was then used to generate an independent set of markers (markers within 10000 kilobases of an index variant were pruned if r^2>0.1), and regions of high LD such as the *HLA* region were also removed. The first 10 principal components were generated from this independent set of markers. Plots of the principal components were used to assess the quality of the matching.

3.3.4 Exploring genetic architecture

After the construction of the nsCL/P meta-analysis summary statistics and the matched case-control sample, several different methods were used to explore the genetic architecture of nsCL/P. These methods included; the PTDT, LD score regression, GCTA and AVENGEME.

3.3.4.1 Polygenic transmission disequilibrium test (PTDT)

The PTDT uses parent-offspring trios to infer the polygenic architecture of a trait or polygenic overlap between two independent traits. The PTDT detects over-transmission of polygenic risk for a trait from unaffected parents to an affected child by constructing PRS from a training sample in an independent target sample of parent-offspring trios [158].

PRS were derived using summary statistics from the Bonn-II study [23] at a range of P-value inclusion thresholds from P<0.000001 to P≤1, i.e. including all SNPs in the summary statistics, (Thresholds: 0.000001, 0.000005, 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1), and then constructed in the imputed ICC nsCL/P European and Asian pedigrees, separately. The PRS included independent SNPs generated by LD clumping the summary statistics in PLINK (clumps were formed of an index variant and other SNPs within 250kb of the index variant with r^2>0.1), using the relevant ICC pedigree (Asian or European) as the reference panel for LD. The standardised difference in PRS between the unaffected parental controls and the affected children was calculated and the PRS P-value inclusion threshold with the strongest evidence of over-transmission was determined using a t-test.

A stratified PTDT was used to investigate the genetic architecture of nsCL/P, i.e. which allele frequency range contributes most to the aetiology of nsCL/P. It was decided to test this on all SNPs (i.e. the PRS with a P-value inclusion threshold of 1) and only in the European pedigrees. The 1000 Genomes (Phase 3) reference panel [163], was used to stratify the LD pruned PRS, into 5 MAF bins (“0<x≤0.1”, “0.1<x≤0.2”, “0.2<x≤0.3”, “0.3<x≤0.4”, “0.4<x≤0.5”). The PTDT was then run separately on each of the 5 strata.

3.3.4.2 Sibling recurrence rate

A simplistic method of estimating heritability of a dichotomous trait, is to use the population prevalence and familial recurrence rate of a trait to estimate the narrow-sense heritability estimate H_2. If we assume a population prevalence of K_P and a sibling recurrence risk of K_{SIB}, the contribution of additive genotypic variation across the population (represented by Σ) can be estimated as a function of the

difference between the sibling recurrence risk and the population prevalence: $\Sigma = 2K_P(K_{SIB}\text{-}K_P)$ while the phenotypic variation across the population (represented by V) can be modelled as a function of the population prevalence: V=K_P(1-K_P). H_2 can then be approximated by the variance in genotype divided by the variance in phenotype: $H_2 = \Sigma$/V. In this instance, sibling recurrence risk and population prevalence estimates for nsCL/P were taken from previous publications [244,245].

3.3.4.3 Linkage Disequilibrium (LD) score regression

LD between a variant of interest and a causal variant causes inflation of the test-statistic proportion to the degree of LD with the causal variant [225]. It has been demonstrated that the SNP heritability of a trait can be estimated by a regression of the test statistic against LD-scores, which are generated from reference panels such as 1000 Genomes [163] and measure the amount of variation tagged by each variant in the genome for each variant [225]. LD-score regression has further applications such as partitioning heritability into functional categories [246] and estimating genetic correlation [247]. Advantages of LD-score regression are that it can be applied to summary data and that the method is automated in a curated web-interface [248]. The main disadvantage of LD-score regression is that the method has lower statistical power compared to methods that use individual level data.

The python package LDSC.py was used to run LD score regression on the summary statistics of the ICC TDT, the Bonn-II Study and the combined TDT/Bonn-II meta-analysis, using LD score files from the 1000 Genomes (Phase 3) CEU data. Observational heritability estimates were converted to the liability scale using a population prevalence from a previous publication [249] and the sample prevalence calculated from the data.

The input parameters for LD score regression are the Z scores and sample sizes for each SNP, which is used to weight the contribution of each SNP to the heritability estimate. Sample sizes for each SNP are trivial to calculate for a case-control design, but non-trivial for a TDT design because the TDT requires at least one parent to be heterozygous for the over-transmission of a SNP to be tested for that parent-offspring trio. Therefore, I tested the effect of changing the sample size parameter for the TDT study on the SNP heritability estimates. As the primary analysis, the sample size for each SNP was considered to be a function of the total number of transmitted and untransmitted alleles (each transmitted or untransmitted allele was assumed to be equivalent to 1 case and 1 control). As a comparative secondary analysis, the sample size for each SNP was considered to be the total number of individuals in the ICC sample (i.e. the number of offspring plus the number of parents).

3.3.4.4 Genome-wide Complex Trait Analysis (GCTA)

The first method proposed for estimating the SNP heritability was Genome-wide Complex Trait Analysis (GCTA) [220]. Genetic relationship matrices (GRM) are constructed that estimate the pairwise relatedness between individuals. The SNP heritability is then estimated using the GRM, LMM and restricted estimated maximum likelihood (REML). The underlying premise is that if individuals who are phenotypically similar are also genotypically similar, then the trait is likely to be heritable. Close relatives such as parents and siblings should be removed before running GCTA analysis because of the large genetic overlap, strong phenotypic similarities and shared environment of closely related individuals [220]. An extension of GCTA, using a bivariate linear mixed model, can be used to estimate the genetic covariance between two traits [250].

GCTA is widely used method for estimating SNP heritability when individual level data are available. However, there are several published criticisms of the method [222,223,251,252]. The primary criticism is that the underlying model used in GCTA makes many strong assumptions that do not necessarily hold. These assumptions pertain to the relationship between the expected amount of heritability assigned to each SNP with respect to MAF and LD [222].

The admixture-matched nsCL/P case-UK Biobank control sample was used to estimate the SNP heritability of nsCL/P using GCTA. First, GCTA was run on the initial matched sample adjusting for the first 10 principal components. A population prevalence, taken from a previous publication [245] and the sample prevalence (calculated from the data) were used to convert the observed scale heritability estimates to the liability scale. Second, GCTA was rerun after removing related individuals (>0.025 on the GRM). Third, the effect of removing ancestral outliers and poorly matched individuals on the heritability estimates was investigated, which were determined using arbitrary cut-offs from principal component plots.

3.3.4.5 Additive Variance Explained and Number of Genetic Effects Method of Estimation (AVENGEME)

AVENGEME [224] is a maximum likelihood method for estimating SNP heritability, genetic covariance and the proportion of null SNPs across the genome. These parameters are estimated using the relationship between different inclusion thresholds for PRS and the test statistic for association between phenotype. Scores are generated using summary data from a training sample and are constructed in an independent target sample of unrelated individuals [136,224].

PRS were derived using summary statistics from the Bonn-II study [23] at a range of 20 P-value inclusion thresholds from 0.000001 to 1 (0.000001, 0.000005,

0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 1) and then constructed in the admixture matched nsCL/P case-UK Biobank control sample. The PRS were generated by LD clumping the summary statistics in PLINK (clumps were formed of an index variant and other SNPs within 250kb of the index variant with r^2>0.1), using the admixture matched sample as the reference panel for LD.

A logistic regression was then performed with case-control status as the binary outcome and the Bonn-II PRS as the explanatory variable. Analysis was run both unadjusted and after adjusting for the first 10 principal components. The Z scores from the logistic regression and other relevant parameters were used to estimate the SNP heritability using AVENGEME. Parameters in the calculation included: sample prevalences, population prevalence and the number of SNPs common to both samples. The sample prevalences and number of SNPs common to both samples were derived from the data while the population prevalence was taken from a previous publication [245]. It was assumed that the genetic architectures in the two samples are identical, i.e. the phenotyping of nsCL/P and ancestry are consistent across the two datasets.

3.3.5 AVENGEME simulations

The sensitivity of the AVENGEME heritability estimates to batch-related systematic differences between cases and controls was tested. It was assumed that if the PRS themselves are not associated with the genotypic batch differences, then the heritability estimate may be unbiased.

The AVENGEME analysis was repeated in the admixture matched sample, with the Bonn-II summary statistics randomly rearranged (each SNP was randomly

allocated a P-value and effect size from another SNP). If AVENGEME heritability estimates are biased by poor quality case-control matching, then randomly arranged summary statistics should generate a heritability estimate significantly distinct from 0.

To test this, 100 simulations were run using AVENGEME on the rearranged Bonn-II summary statistics and the admixture matched case-control sample. It is important to note that in this instance, assuming; substantial standard error in each heritability estimate, heritability estimates constrained between 0 and 1, and no true genetic effect, one would expect the mean of the simulations to be non-zero. This is because if there is no true effect, approximately half of the PRS in the simulations will be positively associated with case-control status while the other half will be negatively associated. However, heritability estimates have a lower bound of 0, meaning the mean of all simulations will be greater than 0.

3.4 Results

3.4.1 Generation of nsCL/P meta-analysis summary statistics

After performing the TDT on the European nsCL/P trios, I meta-analysed the results with the summary statistics from the Bonn-II GWAS. Although there were some slight differences, the P-values were highly concordant with the results of the previously published meta-analysis GWAS [24] which used the same data-sets with slightly different QC and analytical methodology (**Table 4**).

Table 4: Comparison of Ludwig et al with meta-analysis summary statistics

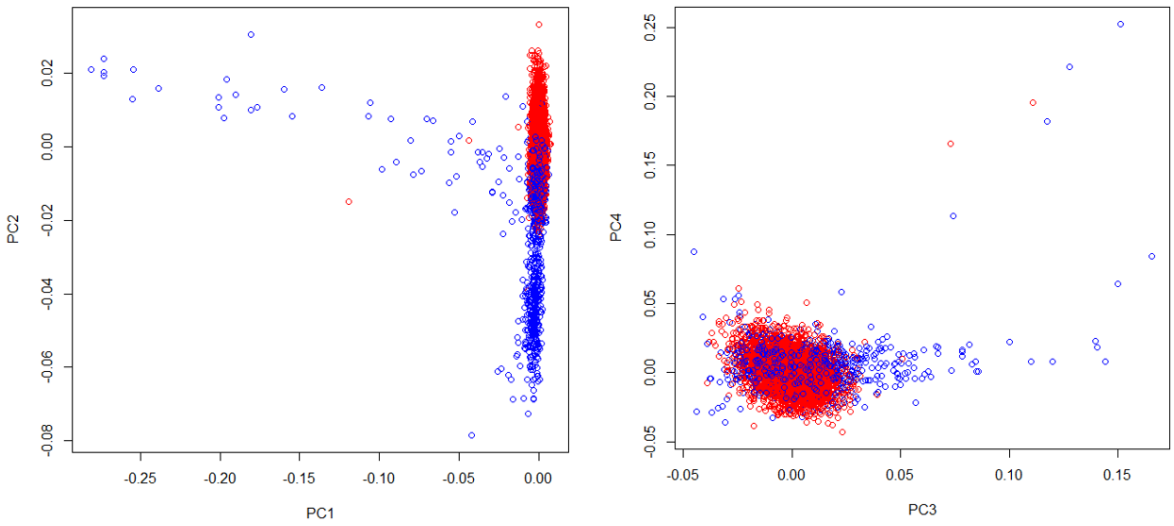
SNP	**(Ludwig et al 2012) European only** **P-value**	**TDT + Bonn-II Meta-Analysis** **P-value**
rs560426	1.02×10^{-6}	4.43×10^{-5}
rs861020	1.78×10^{-6}	1.38×10^{-5}
rs987525	3.94×10^{-34}	7.95×10^{-20} [1]
rs7078160	2.81×10^{-8}	3.99×10^{-7}
rs227731	4.26×10^{-8}	4.50×10^{-7}
rs13041247	7.41×10^{-4}	2.25×10^{-3}
rs742071	2.63×10^{-7}	4.07×10^{-6}
rs7590268	4.05×10^{-8}	2.17×10^{-6}
rs7632427	4.20×10^{-5}	2.62×10^{-4}
rs12543318	1.02×10^{-6}	1.49×10^{-5}
rs8001641	6.20×10^{-10}	4.41×10^{-8}
rs1873147	2.81×10^{-8}	4.22×10^{-7}

[1] rs987525 was removed in the trios in QC hence the discrepancy in P values

3.4.2 Case-control matching

Admixture components were derived in the joint sample of UK Biobank controls and nsCL/P cases. Propensity score matching was then used to construct a matched case-control sample. However, in the matched sample, evidence was found of large systematic genome-wide differences between the nsCL/P cases and the matched UK Biobank controls. Principal component plots suggested that the second principal component divides cases from controls and demonstrated the substantial ancestral heterogeneity across the nsCL/P cases (**Figure 5**). A regression of principal components on case-control status suggested systematic differences between cases and controls, which extended across many of the first 10 principal components (**Table 5**).

Figure 5: Principal component plots of admixture matched sample



Key: Blue: nsCL/P cases and Red: Matched UK Biobank controls

Table 5: Association of first 10 Principal Components with case-control status in matched sample

Principal Component	**P-value**
PC1	0.0001
PC2	$<1.0x10^{-15}$
PC3	$<1.0x10^{-15}$
PC4	$<1.0x10^{-15}$
PC5	1.07^{-13}
PC6	$<1.0x10^{-15}$
PC7	0.040
PC8	0.044
PC9	$2.43x10^{-7}$
PC10	0.64

3.4.3 Polygenic transmission disequilibrium test (PTDT)

nsCL/P PRS from the Bonn-II study were constructed in the ICC European and Asian nsCL/P parent-offspring trios. The PTDT demonstrated consistent over-transmission of nsCL/P genetic risk scores across all inclusion thresholds in European trios. The strongest association was observed when using 17 SNPs with P < 0.00001; affected offspring had 0.32 (95% C.I. 0.25, 0.39) S.D. higher nsCL/P polygenic risk score than their unaffected parents ($P = 3.5 \times 10^{-18}$) (**Table 6**).

In the Asian trios, there was similar evidence of over-transmission, but the magnitude was smaller than in Europeans and strong association was not observed across all inclusion thresholds. As in Europeans, the strongest association was observed using a P value threshold of 0.00001; affected offspring had 0.19 (95% C.I. 0.12, 0.25) S.D. higher nsCL/P polygenic risk score than their unaffected parents ($P = 1.7 \times 10^{-7}$) (**Table 6**).

Table 6: Polygenic Transmission of nsCL/P genetic risk variants in independent European and Asian trios

P-value Inclusion Threshold	**European Trios (N = 2209)**			**Asian Trios (N = 2593)**		
	Number of SNPs	**Beta (95% C.I.)[1]**	**P Value**	**Number of SNPs**	**Beta (95% C.I.)[1]**	**P Value**
0.000001	4	0.30 (0.23, 0.37)	$1.8x\ 10^{-16}$	10	0.12 (0.05, 0.19)	$8.0x\ 10^{-4}$
0.000005	13	0.30 (0.23, 0.37)	$1.1x\ 10^{-16}$	18	0.16 (0.09, 0.23)	$7.0x\ 10^{-6}$
0.00001	**17**	**0.32 (0.25, 0.39)**	**$3.5x\ 10^{-18}$**	**23**	**0.19 (0.12, 0.25)**	**$1.7x\ 10^{-7}$**
0.00005	44	0.25 (0.18, 0.33)	$1.0x\ 10^{-11}$	53	0.16 (0.09, 0.13)	$1.1x\ 10^{-5}$
0.0001	69	0.24 (0.17, 0.31)	$2.1x\ 10^{-11}$	81	0.12 (0.05, 0.19)	0.001
0.0005	222	0.17 (0.10, 0.23)	$2.7x\ 10^{-6}$	244	0.06 (0.00, 0.13)	0.056
0.001	405	0.15 (0.08, 0.22)	$4.6x\ 10^{-5}$	437	0.08 (0.01, 0.14)	0.020
0.005	1,626	0.19 (0.12, 0.26)	$5.3x\ 10^{-8}$	1793	0.12 (0.05, 0.18)	$3.5x10^{-4}$
0.01	3,002	0.16 (0.09, 0.23)	$9.2x\ 10^{-6}$	3,334	0.10 (0.04, 0.17)	0.002
0.05	11,400	0.16 (0.10, 0.23)	$1.7x\ 10^{-6}$	13,222	0.09 (0.02, 0.15)	0.009
0.1	20,133	0.16 (0.10, 0.23)	$2.2x\ 10^{-6}$	23,421	0.06 (-0.01, 0.12)	0.096
0.5	64,727	0.17 (0.10, 0.24)	$1.1x\ 10^{-6}$	74,832	0.03 (-0.03, 0.10)	0.31
1	92,527	0.16 (0.09, 0.23)	$2.6x\ 10^{-6}$	107,809	0.03 (-0.04, 0.10)	0.38

1 Standardised difference in genetic score between parents and offspring

After stratifying SNPs by MAF, PTDT results suggested that high frequency nsCL/P variants on the genotyping chip may be over-transmitted more than lower frequency variants. The largest magnitude of over-transmission was observed for SNPs with a MAF greater than 0.4; Beta = 0.16 (95% C.I. 0.09, 0.23, P value = 0.00002) with the lowest over-transmission observed for SNPs with a MAF less than 0.1; Beta = 0.05 (95% C.I. -0.02, 0.13, P value = 0.16) (**Table 7**).

Table 7: Polygenic Transmission of SNPs by minor allele frequency

MAF Inclusion Threshold	Number of SNPs	Beta (95% C.I.)[1]	P Value
0.0 < x < 0.1	29,034	0.05 (-0.02, 0.13)	0.16
0.1 < x < 0.2	25,592	0.07 (0.01, 0.14)	0.03
0.2 < x < 0.3	15,516	0.11 (0.04, 0.18)	0.001
0.3 < x < 0.4	11,801	0.06 (-0.01, 0.13)	0.10
0.4 < x < 0.5	10,313	0.16 (0.09, 0.23)	0.00002

1 Standardised difference in genetic score between parents and offspring

3.4.4 Narrow-sense heritability using sibling recurrence risk

The narrow-sense heritability estimate of nsCL/P was crudely estimated using population prevalence and sibling recurrence estimates. The estimated population prevalence of nsCL/P is around 1 in a 1000 or 0.1% [249]. The relative risk of sibling recurrence was estimated to be 32 in a Norwegian medical birth registry [244]. Assuming a population prevalence of 0.1%, the sibling recurrence risk estimate is 3.2%. Assuming these parameters and the equation detailed earlier in **Chapter 3.3.4.2**, the narrow sense heritability estimate of nsCL/P is 6.2%.

3.4.5 Linkage disequilibrium score regression

In the primary analysis, (the sample size of each SNP was assumed to be a function of the number of untransmitted and transmitted alleles) the LD score regression SNP heritability estimate of the combined meta-analysis was 0.33 (95% C.I. 0.14, 0.51). However, there was much greater uncertainty in the heritability

estimates for the ICC-TDT and Bonn-II summary statistics, respectively 0.17 (95% C.I. 0, 0.62) and 0.19 (95% C.I. 0, 0.51) likely reflecting the smaller sample sizes (**Table 8**).

Table 8: Linkage Disequilibrium Score Regression results of two independent samples and the combined meta-analysis

Sample	**Sample**	**Sample size parameter**	**Liability scale h^2 (95% C.I.)**
ICC TDT	808 cases and 1462 parental controls	Total sample size[1]	0.04 [0, 0.23)
		2(U+T)[2]	**0.17 [0, 0.62)**
Bonn-II study	401 cases and 1301 controls	**Total sample size[1]**	**0.19 [0. 0.51)**
Combined Meta-analysis	1209 cases and 2763 related and unrelated controls	Total sample size[1]	0.20 (0.08, 0.32)
		2(U+T)[2]	**0.33 (0.14, 0.51)**

1 Total number of cases and controls

2 In the TDT study, the sample size for each SNP was assumed to be double the number of U (untransmitted alleles) and T (transmitted alleles)

The sample size parameter for the contribution of each SNP affected SNP heritability estimates. Although confidence intervals overlapped, the two methods of calculating the sample size parameter gave somewhat discordant effect estimates for both the ICC-TDT sample alone (0.04 and 0.17), and the meta-analysis summary statistics (0.20 and 0.33). Not accounting for the number of homozygous parents for each SNP resulted in a lower SNP heritability effect estimate. This reduction could be because lower MAF SNPs, where over-transmission was only tested within a small sample of trios, are given the same weighting as higher frequency SNPs.

3.4.6 Genome-wide Complex Trait Analysis

GCTA was used to estimate the SNP heritability of nsCL/P using the matched 838 nsCL/P cases and 3352 UK Biobank controls sample. The SNP heritability estimates for nsCL/P from GCTA on the observed scale were 0.79-0.91 compared to

the liability scale SNP heritability estimates of 0.45-0.70. Removing poorly matched individuals and ancestral outliers, using the principal component plots, increased the liability heritability estimates, possibly because estimates are sensitive to sample prevalence, which decreased as outliers removed were predominantly cases (**Table 9**). GCTA on a case-control sample is sensitive to sub-structure differences between cases and controls which have been shown to exist, and the inflated heritability estimates reflect these differences. It is worth noting that ordinarily removing cases or controls without removing their respective matches may induce bias because it ignores the case-control matching. However, given the poor quality of the matching in this instance (see **Figure 5** shown previously), this is unlikely to have a negative effect.

Table 9: GCTA estimates from Admixture-Matched sample with outlier removal

Sample Size	**Heritability Estimate (95% C.I.)**		**Description**
	Observed Scale	**Liability Scale**	
838 Cases and 3352 Controls	0.91 (0.79, 1]	0.50 (0.43, 0.57)	4 Biobank controls matched to each case
792 Cases and 3313 Controls	0.79 (0.65, 0.93)	0.45 (0.37, 0.52)	Relateds removed (>0.025 from GRM)
684 Cases and 3281 Controls	0.80 (0.66, 0.94)	0.49 (0.40, 0.58)	Extreme ancestral outliers removed
446 Cases and 3281 Controls	0.84 (0.69, 0.99)	0.70 (0.57, 0.82)	Largely unmatched cases removed

3.4.7 AVENGEME heritability and simulation results

AVENGEME was used to estimate the SNP heritability of nsCL/P using the matched nsCL/P cases and UK Biobank controls sample. PRS from the Bonn-II study, at different inclusion thresholds, were constructed in the matched sample. The SNP heritability estimates for nsCL/P from AVENGEME were 0.20 (95% C.I. 0.18,

0.22) unadjusted and 0.16 (95% C.I. 0.14, 0.18) when adjusted for the first 10 principal components.

Running simulations of randomly rearranged nsCL/P summary statistics showed that batch and ancestry differences have a minor effect on SNP heritability estimates using AVENGEME because the average estimates across simulations were very close to 0 despite being constrained between 0 and 1. Additionally the results suggested that adjusting for principal components may not necessarily improve accuracy; the median heritability estimate of the simulated data was 0.000 (95% C.I., 0.000, 0.018) without adjusting for principal components and 0.000 (95% C.I., 0.000, 0.027) when adjusting (**Table 10)**.

Table 10: AVENGEME heritability estimates when using random effect sizes from 100 simulations

	Heritability estimate (h2)	
	Unadjusted	**Adjusted for 10 PC**
Mean (95% C.I.)	0.017 (0.008, 0.036)	0.015 (0.006, 0.036)
Median (95% C.I.)	0.000 (0.000, 0.018)	0.000 (0.000, 0.027)
Maximum (95% C.I.)	0.139 (0.107, 0.173)	0.106 (0.078, 0.136)

3.5 Discussion

In this chapter, individual level genotypes and pedigree data were used to explore the genetic architecture of nsCL/P and estimate the proportion of heritability explained by common genetic variation captured by a genotyping chip. Triangulating the results from several different methods, strong evidence was found for a substantial role of common genetic variation in the aetiology of nsCL/P and a highly polygenic genetic architecture. These findings have several important implications for work in later chapters. Firstly, the evidence for polygenicity suggests that nsCL/P PRS, which are used in later chapters to test genetic overlap, can be used effectively

as genetic proxies for liability to nsCL/P. Secondly, the high effect sizes and allele frequencies of the known risk loci suggest that MR analyses with liability to nsCL/P as the exposure are likely to be well-powered. Finally, a polygenic architecture for nsCL/P is consistent with the proposed multifactorial liability model of inheritance.

The PTDT showed consistent over-transmission of nsCL/P polygenic risk from unaffected parents to affected offspring across all P value thresholds in European trios and over some thresholds in Asian trios, supporting the notion that nsCL/P has a polygenic component driven by common genetic variation. The observation of sub-clinical craniofacial phenotypes in individuals with nsCL/P and unaffected relatives, such as lip pits [93], orbicularis oris muscle defects [96,97] and dental anomalies [102], is consistent with a polygenic architecture for nsCL/P. The most predictive score in the PTDT including less than 20 SNPs, suggests that genetic risk for nsCL/P may be largely attributable to a modest number of common variants, consistent with the previous finding that the majority of the SNP heritability in Europeans is explained by 24 known risk loci [26]. A further important finding is that nsCL/P polygenic risk scores generated in Europeans are still predictive in Asian populations, albeit to a lesser extent, which is consistent with the difference in variation explained by known risk loci in European and Asian populations [26,28]. MAF stratified PTDT analysis further supported the importance of high frequency genetic variation, although the extent to which lower frequency variation is tagged by SNPs on the genotyping chip is largely unclear.

A modest narrow-sense heritability estimate of 6.2% using sibling recurrence risk and prevalence may be attributable to both the crudeness of the method and to the low population prevalence of nsCL/P (the magnitude of the denominator in the calculation is inversely proportional to the prevalence). LD score regression and

AVENGEME SNP heritability estimates of 33% and 20% respectively, further support the role of common variation, and are consistent with a previous SNP heritability estimate of 30% [26].

Despite the efforts to match nsCL/P cases to controls from the UK Biobank on ancestry, there was strong evidence of systematic differences between cases and matched controls. These systematic differences could be attributable to the difficulties matching the ancestrally heterogeneous ICC nsCL/P cases, sampled from across the US and Scandinavia, to individuals in the UK Biobank. Alternatively, the differences may be caused by batch effects, i.e. differences in genotyping chips and related processes. The GCTA estimates of SNP heritability of 45-70% on the liability scale are higher than the estimates from other methods in this chapter and previous estimates [26], but are likely inflated by the systematic differences in the matched sample. Contrastingly, simulations showed that any systematic differences in the matched sample are unlikely to have biased the AVENGEME heritability estimates.

The findings of strong evidence for nsCL/P having a polygenic architecture and a substantial SNP heritability are highly concordant with previous findings. Previous estimates of the narrow-sense heritability of nsCL/P from twin and pedigree studies ranged between 40 and 90% [198,199], which are higher than our SNP heritability estimates, but this may be explained by narrow-sense estimates including the effects of rare variation. As discussed previously, a published SNP heritability estimate is also highly concordant with our findings [26].

The major strength of the analysis is the thoroughness of evaluating many different heritability estimation methods (narrow-sense and SNP) for a complex trait where the genetic architecture is largely unknown. A further strength is the use of

both European and Asian trios which allowed for the evaluation of the effect of population differences in polygenic scoring methods. The analysis also featured some relatively novel approaches in exploring genetic architecture such as stratification on allele frequency in the PTDT. Finally, the systematic differences in the matched sample, although a limitation in most regards, revealed that methods using PRS such as AVENGEME may be largely unaffected by genotyping chip differences.

The major limitation of the study is that the matched case-control samples used were not homogeneous; this had a sizeable effect on the heritability estimates from GCTA which are therefore likely unreliable. Despite the efforts to match cases to controls on ancestry, genotyping chip differences between cases and controls or genuine ancestral differences resulted in population substructure differences between the cases and controls in the matched samples. Although some nsCL/P cases mapped relatively well with the matched controls on principal component plots, the plots suggested that systematic differences between cases and controls were highly prevalent even after the removal of ancestral outliers. A possible reason for this is the heterogeneity between the European nsCL/P cases; the cases are sampled from different populations including Denmark, Norway and the USA. Another limitation is that of statistical power, the number of nsCL/P cases used in the analyses was relatively low; LD score regression recommends sample sizes greater than 3000 and this was reflected by large confidence intervals. Statistical power may be lowered further by phenotypic heterogeneity within the nsCL/P phenotype; there is increasing evidence that nsCLP and nsCLO may have distinct aetiologies [8,13,27,253]. Considering these phenotypic differences, treating nsCL/P as a homogenous

phenotype in our analyses may have resulted in measurement error, as the genetic architecture and SNP heritability may differ between the nsCL/P subtypes.

To conclude; the PTDT demonstrated that nsCL/P is a highly polygenic trait and AVENGEME / LD score regression analyses estimated that common genetic variation explains between 20 and 33% of the phenotypic variance in nsCL/P. These findings, in conjunction with previous findings, imply that both PRS and MR can be utilised effectively in an epidemiological context to explore causal relationships involving liability to nsCL/P. AVENGEME has been shown in this instance to give heritability estimates largely unaffected by batch and could be the most appropriate heritability method to use in samples affected by batch. The difficulties with matching cases to controls on ancestry across samples with different genotyping chips was also shown to be a non-trivial undertaking and should be considered carefully, especially for designs planning to use GCTA.

Chapter 4: Epigenetics and nsCL/P

4.1 Abstract

Many nsCL/P genetic risk variants identified in GWAS reside in non-protein-coding regions with an unclear function. One possibility is that genetic risk variants influence susceptibility to nsCL/P through gene expression pathways, such as those involving DNA methylation.

MR and joint likelihood mapping were used to identify putative loci where genetic liability to nsCL/P may be mediated by variation in DNA methylation, using nsCL/P GWAS summary data and methylation data from four studies. The primary analyses used DNA methylation in blood, so the correlation between DNA methylation in blood and more appropriate tissues (lip/palate) was estimated for relevant CpG sites.

Evidence was found at three independent loci, *VAX1* (10q25.3)*, LOC146880* (17q23.3) and *NTN1* (17p13.1), that liability to nsCL/P and variation in DNA methylation might be driven by the same genetic variant. Follow up analyses using DNA methylation data, derived from lip and palate tissue, and gene expression catalogues provided further insight into possible biological mechanisms. Genetic variation may increase liability to nsCL/P by influencing DNA methylation and gene expression at *VAX1, LOC146880* and *NTN1*.

4.2 Introduction

4.2.1 Epigenetics, gene expression and nsCL/P

GWAS have identified around 40 distinct genetic risk variants for nsCL/P in European and Asian populations [21-24,28-30,218,254] but many variants reside in non-

protein coding regions and so their functional relevance remains unclear. One possibility is that genetic risk variants for nsCL/P may act through gene regulation pathways. For example, a previous study demonstrated that a major nsCL/P risk locus, a non-coding interval at 8q24, regulates gene expression in the developing murine face, suggesting similar mechanisms in humans [32].

Epigenetics refers to mitotically (and perhaps, controversially meiotically) heritable changes in gene expression that are not explained by changes to the DNA sequence [255]. Epigenetic processes, which include DNA methylation, histone modification and non-coding RNAs, can effect gene expression by influencing transcription [256]. There is increasing evidence that epigenetic mechanisms, such as DNA methylation, play a role in the development of OFCs [33-36], potentially via changes to gene expression.

4.2.2 DNA methylation

DNA methylation, the most widely-studied epigenetic mechanism is a reversible modification of DNA, typically involving the addition of a methyl group to a cytosine base in the base-pair sequence CpG. The CpG pairing has been shown to have a lower than expected frequency throughout the genome but occurs more frequently in regions known as CpG islands [256]. CpG islands are found in the coding regions of around half of all transcribed genes and are thought to play a major role in gene expression [256]. DNA methylation can influence gene expression by silencing transcription; one possible mechanism is that methylation can block interactions between the DNA and proteins involved in transcription [257,258].

DNA methylation has been proposed as a potential mediator of environmental effects on disease risk because it can be altered by exogenous stimuli and these alterations could affect gene expression [259]. For example, prenatal tobacco smoke

exposure causes changes in methylation [260] and there is some evidence that these changes may mediate the effects of maternal smoking on reduced birthweight in the offspring [261].

Two groups of diseases where DNA methylation has been shown to be highly relevant are X-linked Mendelian syndromes and cancer. The importance of DNA methylation in development is illustrated by the effect of mutations in the *MECP2* gene, which codes for a protein involved in the binding of methylated DNA. Mutations in *MECP2* can cause Rett syndrome, Angelman syndrome and non-specific X-linked mental retardation [262]. An aetiological link between methylation and cancer is suggested by the differential methylation of malignant tumour cells compared to healthy cells [263].

4.2.3 Investigating the role of DNA methylation

The association between DNA methylation at CpG sites across the genome, and complex traits can be investigated using EWAS. However, the direction of effect between DNA methylation and a phenotype is difficult to determine, as epigenetic marks are susceptible to both confounding and reverse-causation. Unlike germline DNA, methylation can vary across the life course and so can be affected by the presence of disease as well as many potential common confounders from observational epidemiological studies, such as an individual's age [264,265].

Further potential confounders when investigating a possible causal role of DNA methylation are batch effects [165] and cellular heterogeneity [178]. Batch effects between different samples can be caused by variation in chips, pre-processing methods or assaying. Cell counts are another potential confounder as DNA methylation varies greatly across different cell types and cellular heterogeneity has been shown to account for a large proportion of epigenetic differences [178,265]. In a

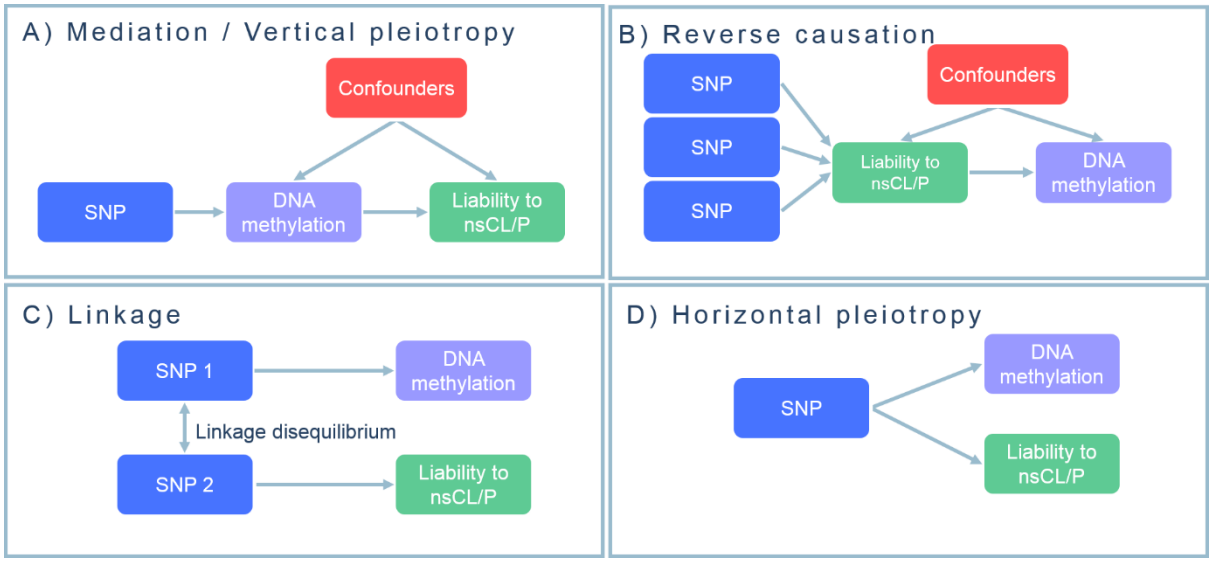
similar vein to cellular heterogeneity, methylation can vary greatly between different tissue types and it is therefore important to measure methylation in an appropriate tissue [92].

4.2.4 Genetics of DNA methylation and epigenetic MR

A substantial proportion of the variation in DNA methylation is explained by variation in the germline genome, such variants are referred to as methylation quantitative trait loci (mQTL). mQTL can affect methylation in cis, i.e. affect the methylation of nearby areas, or in trans, affecting sites further away [266].

mQTL can be used as instruments for DNA methylation in an MR framework to more formally investigate the relationship between DNA methylation and a phenotype [259]. If a SNP, which is a mQTL, is also associated with nsCL/P, there are several distinct possibilities. The first possibility is that the effect of the SNP on nsCL/P is mediated by DNA methylation, i.e. the SNP affects DNA methylation which then in turn affects the phenotype (vertical or mediatory pleiotropy). The second possibility is that the SNP affects the phenotype through a pathway independent of DNA methylation, and liability to nsCL/P causes differences in methylation. In some instances, this reverse direction could be thought of as reverse-causation; for example, if atherosclerosis (often a precursor to more serious heart disease symptoms) causes changes in methylation. However, in this instance the interpretation of liability to nsCL/P causing differences in methylation as reverse causation is less clear. The third possibility is that the causal SNP influencing methylation at a probe is in strong LD with another SNP that independently affects the risk of nsCL/P (linkage disequilibrium). The fourth possibility is that nsCL/P and DNA methylation are influenced by the same SNP, but via independent pathways (horizontal pleiotropy). See **Figure 6**.

Figure 6: Possible explanations for an association between mQTL and nsCL/P. In this chapter, I attempted to identify loci where genetic influences on nsCL/P are mediated by DNA methylation, i.e. the top left-hand box.



The principles of MR and joint likelihood co-localization can be used to determine the most likely of the four possibilities at each locus. The main advantage of using an MR approach is that the germline genome is fixed from birth, alleviating the possibility of reverse-causation, and may also be less susceptible to confounding. However, one limitation is that it is currently difficult to differentiate between mediation and horizontal pleiotropy because the majority of CpGs are instrumented by a single genetic variant [259,267,268].

4.2.5 Chapter aims

In this chapter, genetic and epigenetic data from several different cohorts were used to investigate if DNA methylation mediates the effect of genetic risk variants for nsCL/P. The four possibilities, outlined previously, were evaluated using bidirectional MR and joint-likelihood co-localisation. It was not possible to distinguish between mediation and horizontal pleiotropy so inferred mediating CpGs are putative.

mQTL relevant to putative mediating CpGs were investigated with regards to overlap with gene expression (a potential functional mechanism) using publicly available databases. Putative CpGs were compared with the results of a previous nsCL/P EWAS, derived from whole blood in a Brazilian population, to determine if they were differentially methylated between nsCL/P cases and controls.

Given that primary analyses used DNA methylation derived in blood, which might not be representative of the developing orofacial tissues [36], correlations between DNA methylation in blood and lip/palate tissue in the same individuals were estimated. Additionally, considering the growing evidence that different subtypes of OFCs have distinct aetiologies [8], it was investigated whether putative CpGs are differentially methylated in blood samples from children with different OFC subtypes. In this chapter, Dr Gemma Sharp performed analyses that involved the Cleft Collective data and the methWAS cohort. I performed all other described analyses.

4.3 Materials and methods

4.3.1 Study participants

4.3.1.1 ICC and Bonn-II

The nsCL/P meta-analysis GWAS summary statistics derived in **Chapter 3** were used for information on nsCL/P genetic risk variants. In brief, the meta-analysis summary statistics included ICC TDT results from 638 parent-offspring trios and 178 offspring duos of European descent, meta-analysed with GWAS summary results on 399 cases and 1,318 controls from the Bonn-II study.

LiftOver [269] was used to convert the genome positions in the nsCL/P summary statistics to the most recent genome build 37. Finally, PLINK [161] was used to clump the results according to LD (r^2<0.001) using ALSPAC as a reference panel,

within a 250 kb region around each index variant, and to generate a set of independent SNPs for the pipeline.

4.3.1.2 ALSPAC

To identify mQTL (SNPs associated with DNA methylation), data were used from ALSPAC [270,271]. Genome-wide DNA methylation and genotype data are available for ARIES, a subset of ALSPAC [164], which has been described in more detail in **Chapter 2**. These data have previously been used to generate a database of mQTL (http://www.mqtldb.org/) [266]. The database contains summary statistics for all mQTL with a P-value $<1x10^{-7}$ for the association between SNP and CpG.

For the purposes of this study, mQTL identified in cord blood samples collected at birth were most relevant (the closest available time point to the orofacial developmental period). For one analysis (the reverse two-sample MR), specific CpG-SNP associations were required that were unavailable from mQTLdb.org. Therefore, for required CpGs, the methods from the original study were replicated: individuals with missing genotype or covariate data were excluded, leaving 787 children. The methylation data were then rank-normalised to remove outliers and covariates, potential batch effects and the influence of cell heterogeneity were controlled for by regressing data points on sex, the first 10 ancestry principal components, bisulfite-converted DNA batch and blood cell proportions estimated using the Houseman method [178,179]. Residuals were then calculated, which were used as the outcome variable in a linear regression model in PLINK[161] to calculate the relevant CpG-SNP associations.

Finally, any mQTL acting in trans (i.e. any SNP associated with a CpG site more than 1 million base pairs away) and any CpGs that have been flagged as

potentially problematic, (for example, cross-hybridising probes) according to a previous publication, were removed [272].

4.3.1.3 GOYA

mQTL of interest, identified in ALSPAC, were followed up for replication using genotype and cord blood DNA methylation data from the GOYA cohort [177]. More information on GOYA is contained in **Chapter 2**.

Genotype and cord blood DNA methylation data were available for 1000 children. The methods described above for ALSPAC were replicated. Individuals with missing genotype or covariate data were excluded, leaving 889 children. SNPs with missingness (>5%) were also removed using PLINK. As in ALSPAC, the methylation data were rank-normalised to remove outliers and covariates, potential batch effects and the influence of cell heterogeneity were adjusted for by regressing data points on sex, the first 10 ancestry principal components, DNA batch and blood cell proportions estimated using the Houseman method [178,179]. Residuals were then used as the outcome variable in a linear regression model in PLINK to calculate the relevant CpG-SNP associations.

4.3.1.4 Expression quantitative trait loci (eQTLs) databases

To identify eQTL (SNPs associated with gene expression), two gene expression databases were used. The GTEx database [182,183] and the NNC eQTL Catalog [184]. More information on these databases is contained in **Chapter 2**.

4.3.1.5 methWAS cohort

To assess whether methylation at nsCL/P-associated CpGs (identified through MR) differs between nsCL/P cases and controls, relevant CpGs were looked up in a recently-published EWAS [36] which compared blood DNA methylation profiles

in 67 non-familial, nsCL/P cases and 59 age- and sex-matched controls from a Brazilian population. More information on the methWAS cohort is contained in **Chapter 2**.

4.3.1.6 The Cleft Collective

To explore whether methylation at nsCL/P-associated CpGs differs by OFC subtype, the mean methylation values in blood were compared with matched lip/palate tissue samples from 150 children from the United Kingdom, enrolled in the Cleft Collective birth cohort study. More information on the Cleft Collective is contained in **Chapter 2**.

4.3.2 Testing for mediation: Mendelian randomization of the effect of methylation on liability to nsCL/P

The nsCL/P meta-GWAS summary statistics for 543,150 SNPs were LD-pruned (r^2<0.001) to 17,090 independent SNPs using ALSPAC as the reference panel for the LD. These independent SNPs were then merged with 127,215 mQTL from the ALSPAC mQTL database. After removing potentially problematic CpGs and CpGs acting in trans (which may increase the likelihood of horizontal pleiotropy), there were 7,091 independent CpG-SNP pairings for 6,425 distinct CpGs.

The MR-base R package [273] was then used to perform two-sample MR on all CpGs, using mQTL as the exposure variables and nsCL/P as the outcome. In initial analyses, CpGs with one mQTL were tested using the Wald test which is a method of estimating the causal effect when there is only a single valid SNP for the exposure. The causal estimate is estimated as the effect of the SNP on outcome divided by the effect of the SNP on the exposure, and can be interpreted as the unit change in the outcome for a unit change in the exposure [274].

CpGs associated with more than one independent mQTL were tested using the Inverse Variance Weighted (IVW) method, an extension to the Wald Ratio method when there are multiple SNPs. The IVW estimates the causal effect by combining the Wald ratios for each variant in a fixed-effects or random-effects meta-analysis [275]. To account for possible residual LD between mQTL, CpGs with more than one mQTL, were retested adjusting for LD between the SNPs using a likelihood-based method [276]. Pair-wise SNP LD was computed using the CEU and GBR populations from the 1000 Genomes in LDlink [277]. As a sensitivity analysis, replication of the SNP-CpG associations with a Bonferroni-corrected MR P-value <0.05 was attempted in GOYA.

4.3.3 Testing for reverse causation: Mendelian randomization of the effect of liability to nsCL/P on methylation

To assess the possibility that liability to nsCL/P causes changes in DNA methylation, MR-Base was used to conduct the reverse two-sample MR. Six SNPS, previously found to be genome-wide significant in Europeans [24], were used as the exposure, proxying for liability to nsCL/P. mQTL from ALSPAC were used as proxies for the methylation outcomes. The IVW method was used as the primary analysis.

4.3.4 Testing for linkage: joint-likelihood mapping to assess co-localisation

Joint likelihood mapping is a method of testing for co-localisation, where two phenotypes are being affected by the same causal variant in a region of interest rather than two independent causal variants affecting each of the phenotypes separately. The likelihood of co-localisation is estimated by exploring the shape of the association curves over a region. Co-localisation can suggest vertical pleiotropy where a genetic variant affects a phenotype which in turn has a downstream effect on the other phenotype [278].

The Joint Likelihood Mapping (JLIM) package in R (jlim.R) [278] was used to test the possibility that liability to nsCL/P and methylation are driven by the same causal effect in each region of interest. To distinguish between separate causal variants, we set the limit of genetic resolution in terms of r^2 to 0.8 (LD). The CEU data from the 1000 Genomes [163] (Phase 3) was used as the reference dataset for LD. The majority of CpGs were associated with only one independent mQTL, so it was not possible to distinguish between mediation/vertical pleiotropy (top left-hand panel of **Figure 6**) and horizontal pleiotropy (bottom right-hand panel of **Figure 6**).

4.3.5 Comparison with gene expression

The previous steps identified CpGs that potentially mediate the effect of genetic variation on susceptibility to nsCL/P. Further evidence for a functional effect would be provided if implicated mQTL also affect gene expression. Therefore, relevant SNPs were looked up in two eQTL databases (GTEx [183] and NESDA NTR Conditional eQTL Catalog [184]) and the estimated effect sizes and P-values for eQTL in various tissues were noted.

4.3.6 Comparison to methWAS EWAS results

At implicated CpGs, data from the methWAS cohort were used to explore differential methylation between nsCL/P cases and controls from the Brazilian methWAS EWAS study. The estimated direction of effect and P-values obtained using the observational EWAS and MR approaches were compared.

4.3.7 Tissue and cleft-subtype-specific variation

At identified CpGs, data from the Cleft Collective were used to explore 1) the correlation between methylation in blood and methylation at the site of the cleft (lip/palate), and 2) variation in mean methylation according to cleft subtype (CLO,

CPO or CLP). One-way ANOVA was used to compare the mean methylation of subtypes, adjusting for sex and surrogate variables designed to capture technical batch and cell composition effects.

4.4 Results

4.4.1 Testing for mediation: Mendelian randomization of the effect of methylation on liability to nsCL/P

Two sample MR was used to identify CpGs where methylation may mediate genetic liability to nsCL/P by performing an MR analysis of methylation on nsCL/P. mQTL from ALSPAC data were used as genetic instruments for methylation at different CpG sites (exposures) and SNPs from the nsCL/P GWAS meta-analysis summary statistics were used as genetic instruments for liability to nsCL/P (outcome). Evidence was found for an effect of methylation on liability to nsCL/P at 26 CpGs after a Bonferroni correction for 6,425 tests (Bonferroni-corrected P-value $<7.8x10^{-6}$, corresponding to an uncorrected P-value <0.05). Of these 26 CpGs, 20 were instrumented by single mQTL and six were instrumented by two mQTL each. When the six CpGs with two mQTL each were re-tested, accounting for the LD between the SNPs, only one (cg02598441 at *LOC146880*) survived correction for multiple testing. These 21 mQTL were therefore taken forward to the reverse-causation step.

As a sensitivity analysis, all 21 of the ALSPAC mQTL were investigated in data from the GOYA cohort. 17 of the 21 CpG-SNP pairings passed QC and were present in the GOYA data, of which 16 replicated as mQTL in the same direction with P<0.05 (**Table 11**).

Table 11: mQTL replication

SNP (allele 1/allele 2; annotated gene)	**CpG (annotated gene)**	**ALSPAC mQTL:** *Effect size[1], P-value*	**GOYA mQTL:** *Effect size, P-value*	**Replicate with GOYA P<0.05**
rs12057415 (T/C; n/a)	cg09549015 (*F3*)	0.26, $9.4*10^{-9}$	0.008, $1.7*10^{-17}$	Yes
rs12057415 (T/C; n/a)	cg26112574 (n/a)	-0.39, $1.2*10^{-19}$	-0.028, $2.7*10^{-43}$	Yes
rs861020 (A/G; *IRF6*)	cg12766975 (*IRF6*)	0.29, $7.8*10^{-9}$	0.031, $2.7*10^{-20}$	Yes
rs861020 (A/G; *IRF6*)	cg09163369 (*C1orf107*)	-0.60, $6.0*10^{-24}$	-0.032, $9.3*10^{-37}$	Yes
rs861020 (A/G; *IRF6*)	cg23166289 (*C1orf107*)	-0.44, $1.7*10^{-15}$	-0.027, $1.1*10^{-21}$	Yes
rs861020 (A/G; *IRF6*)	cg05527609 (*C1orf107*)	0.34, $2.6*10^{-9}$	-0.007, $6.2*10^{-97}$	Yes
rs4422741 (C/T; n/a)	ch.8.2579072R (n/a)	0.75, $1.3*10^{-49}$	0.030, $2.3*10^{-17}$	Yes
rs4752028 (C/T; *SHTN1*)	cg00750430 (*KIAA1598*)	0.34, $3.5*10^{-11}$	0.017, $2.1*10^{-17}$	Yes
rs4752028 (C/T; *SHTN1*)	cg03968911 (*KIAA1598*)	-0.52, $2.7*10^{-24}$	-0.049, $4.5*10^{-45}$	Yes
rs4752028 (C/T; *SHTN1*)	cg11398452 (*VAX1*)	-0.51, $7.6*10^{-27}$	-0.000, 0.27	No
rs1258763 (C/T; n/a)	cg04870120 (n/a)	0.25, $1.6*10^{-9}$	N/A (didn't have SNP)	N/A
rs1873147 (G/A; n/a)	cg04194852 (*TPM1*)	0.24, $4.3*10^{-8}$	0.001, $1.1*10^{-20}$	Yes
rs8076457 (T/C; *NTN1*)	cg18901140 (n/a)	-0.28, $3.2*10^{-8}$	-0.018, $2.2*10^{-7}$	Yes
rs8076457 (T/C; *NTN1*)	cg19788727 (*NTN1*)	-0.36, $8.6*10^{-13}$	-0.016, $5.0*10^{-15}$	Yes
rs8076457 (T/C; *NTN1*)	cg02481697 (*NTN1*)	-0.41, $1.1*10^{-15}$	-0.061, $2.2*10^{-22}$	Yes
rs8076457 (T/C; *NTN1*)	cg01862363 (*NTN1*)	-0.51, $1.6*10^{-24}$	-0.085, $6.6*10^{-30}$	Yes
rs8076457 (T/C; *NTN1*)	cg16107528 (*NTN1*)	-0.49, $1.9*10^{-26}$	-0.039, $6.8*10^{-27}$	Yes
rs1808191 (C/A; *PLEKHM1P1*)	cg14501219 (*LOC146880*)	-0.33, $1.5*10^{-9}$	-0.014, $1.2*10^{-19}$	Yes
rs1991401 (G/A; *CEP95*)	cg02598441 (*LOC146880*)	0.25, $2.2*10^{-8}$	N/A (didn't have SNP)	N/A
rs1808191 (C/A; *PLEKHM1P1*)		0.83, $3.6*10^{-66}$	0.021, $8.2*10^{-112}$	Yes

rs3746101 (T/G; *MKNK2*)	cg05254098 (*MKNK2*)	-0.46, $5.3*10^{-8}$	N/A (didn't have SNP)	N/A
rs3746101 (T/G; *MKNK2*)	cg17068236 (*MKNK2*)	0.57, $8.7*10^{-11}$	N/A (didn't have SNP)	N/A

[1] ALSPAC regression coefficients are on rank-normalised data

4.4.2 Testing for reverse causation: Mendelian randomization of the effect of genetic liability to nsCL/P on methylation

Next, we tested if the association between the mQTL and liability to nsCL/P arose because liability to nsCL/P, a latent measure of nsCL/P, affects methylation. Two sample MR was used to test this possibility, with liability to nsCL/P as the exposure and methylation as the outcome. No evidence was found for liability to nsCL/P influencing variation in methylation at the 21 CpGs (**Table 12**). However, it should be noted at this point that this step is very likely to be limited by statistical power.

4.4.3 Testing for linkage: joint-likelihood mapping to assess co-localisation

Next, a co-localisation method was used to assess if there was evidence that methylation and liability to nsCL/P are driven by the same causal effect at each locus. Of the 20 CpGs instrumented by one mQTL each, evidence was found for co-localisation at four CpGs (cg11398452, cg01862363, cg02481697 and cg16107528). With the addition of the CpG site associated with two mQTL (cg02598441), evidence was found that methylation at five CpGs are putative mediators of genetic liability to nsCL/P at four different SNPs (**Table 12**).

Of these four SNPs, three were available and tested in the imputed GOYA data (rs807647, rs1808191 and rs4752028). Two of the SNPs (intergenic rs8076457 and rs1808191 near *PLEKHM1P1*) consistently replicated as mQTL in GOYA. The

third SNP, rs4752028, replicated as an mQTL for two out of three CpG sites but did not replicate for the CpG-SNP pairing (rs4752028/cg11398452) where there was evidence of co-localisation (**Tables 11 and 12**).

Table 12: Results of the forward (methylation → nsCL/P) and reverse (nsCL/P → methylation) Mendelian randomisation and the co-localisation analyses in ALSPAC.

SNP (allele 1/allele 2; annotated gene)	**CpG (annotated gene)**	**Forward MR:** *effect size [standard error]; P-value*	**Reverse MR:** *effect size [standard error]; P-value*	**Co-localisation:** *JLIM statistic; P-value by permutation*
rs12057415 (T/C; n/a)	cg09549015 (*F3*)	-1.1 [0.2]; $1.1*10^{-6}$	0.04 [0.07]; 0.59	-8.5; 1
rs12057415 (T/C; n/a)	cg26112574 (n/a)	0.7 [0.1]; $1.1*10^{-6}$	0.00 [0.05]; 1.00	-16.7; 1
rs861020 (A/G; *IRF6*)	cg12766975 (*IRF6*)	1.1 [0.2]; $1.1*10^{-6}$	0.00 [0.04]; 0.99	-11.8; 1
rs861020 (A/G; *IRF6*)	cg09163369 (*C1orf107*)	-0.5 [0.1]; $1.1*10^{-6}$	-0.04 [0.08]; 0.59	-36.1; 1
rs861020 (A/G; *IRF6*)	cg23166289 (*C1orf107*)	-0.7 [0.2]; $1.1*10^{-6}$	-0.01 [0.06]; 0.92	-36.8; 1
rs861020 (A/G; *IRF6*)	cg05527609 (*C1orf107*)	0.9 [0.2]; $1.1*10^{-6}$	-0.06 [0.06]; 0.31	-2.1; 0.69
rs4422741 (C/T; n/a)	ch.8.2579072R (n/a)	0.7 [0.1]; $2.1*10^{-10}$	0.11 [0.06]; 0.08	-2.4; 0.91
rs4752028 (C/T; *SHTN1*)	cg00750430 (*SHTN1*)	1.2 [0.2]; $8.7*10^{-9}$	0.13 [0.11]; 0.25	-16.7; 1
rs4752028 (C/T; *SHTN1*)	cg03968911 (*SHTN1*)	-0.8 [0.1]; $8.7*10^{-9}$	-0.11 [0.19]; 0.58	-32.2; 1
rs4752028 (C/T; *SHTN1*)	cg11398452 (*VAX1*)	-0.8 [0.1]; $8.7*10^{-9}$	-0.11 [0.19]; 0.56	30.2; <0.001
rs1258763 (C/T; n/a)	cg04870120 (n/a)	-1.2 [0.3]; $1.3*10^{-6}$	0.04 [0.05]; 0.38	-68.4; 1
rs1873147 (G/A; n/a)	cg04194852 (*TPM1*)	1.5 [0.3]; $1.5*10^{-8}$	0.09 [0.10]; 0.38	-13.1; 1
rs8076457 (T/C; *NTN1*)	cg18901140 (n/a)	-1.1 [0.2]; $3.0*10^{-7}$	0.02 [0.07]; 0.74	-34.6; 1
rs8076457 (T/C; *NTN1*)	cg19788727 (*NTN1*)	-0.9 [0.2]; $3.0*10^{-7}$	0.03 [0.05]; 0.51	-13.9; 1
rs8076457 (T/C; *NTN1*)	cg02481697 (*NTN1*)	-0.8 [0.2]; $3.0*10^{-7}$	0.01 [0.05]; 0.83	0.65; 0.01
rs8076457 (T/C; *NTN1*)	cg01862363 (*NTN1*)	-0.6 [0.1]; $3.0*10^{-7}$	0.03 [0.09]; 0.78	0.11; 0.016
rs8076457 (T/C; *NTN1*)	cg16107528 (*NTN1*)	-0.7 [0.1]; $3.0*10^{-7}$	-0.01 [0.05]; 0.98	4.3; <0.001

rs1808191 (C/A; *PLEKHM1P1*)	cg14501219 (*LOC146880*)	-1.0 [0.2]; $2.9*10^{-6}$	0.09 [0.05]; 0.051	NA[1]
rs1991401 (G/A; *CEP95*)	cg02598441 (*LOC146880*)	0.4 [0.1]; $4.3*10^{-7}$	-0.04 [0.07]; 0.59	NA[2]
rs1808191 (C/A; *PLEKHM1P1*)				
rs3746101 (T/G; *MKNK2*)	cg05254098 (*MKNK2*)	-1.0 [0.2]; $5.0*10^{-6}$	-0.03[0.05]; 0.54	-37.2; 1
rs3746101 (T/G; *MKNK2*)	cg17068236 (*MKNK2*)	0.8 [0.2]; $5.0*10^{-6}$	-0.02 [0.05]; 0.58	-91.0; 1

[1] This region was too sparsely genotyped to apply the co-localisation analysis
[2]This CpG had two mQTL, so we did not apply the co-localisation analysis

4.4.4 Comparison between methylation and gene expression

In a look-up of the four identified SNPs in the GTex and NESDA NTR Conditional eQTL databases, strong evidence was found for rs4752028 at *SHTN1* being an eQTL for *SHTN1*. There was also strong evidence that both rs1808191 at *PLEKHM1P1* and rs1991401 at *CEP95/DDX5* are eQTL for six nearby genes, including *CEP95* and *DDX5*, which were identified through both databases. There was no strong evidence that the intergenic SNP rs8076457 is associated with gene expression in either database which only included pairings meeting a certain statistical threshold (**Table 13**).

Table 13: Associations with gene expression at identified SNPs in two eQTL databases

SNP (annotated gene)	**CpG (annotated gene)**	**GTex:** *Gene, tissue, effect size, P-value*	**NESDA NTR Conditional eQTL Catalog:** *Gene, tissue, effect size, P-value*
rs4752028 (*SHTN1*)	cg11398452 (*VAX1*)	*SHTN1*, whole blood, 0.35, $1.4*10^{-15}$	*SHTN1*, whole blood, 0.68, $2.7*10^{-159}$
rs8076457 (*NTN1*)	cg01862363 (*NTN1*)	N/A	N/A
	cg02481697 (*NTN1*)		
	cg16107528 (*NTN1*)		
rs1808191 (*PLEKHM1P1*)	cg02598441 (*LOC146880*)	*RP13-104F24.3*, sun exposed skin, -0.31, $1.8*10^{-15}$ *SMURF2*, transformed fibroblasts, 0.28, $3.8*10^{-13}$ *has-mir-6080*, sun exposed skin, -0.29, $1.5*10^{-11}$ *PLEKHM1P,* sun exposed skin, -0.13, $3.2*10^{-5}$	N/A
rs1991401 (*CEP95*)		*DDX5*, whole blood, -0.30, $1.8*10^{-19}$ *CEP95,* whole blood, 0.14, $2.3*10^{-7}$ *MILR1,* whole blood, 0.24, $1.4*10^{-5}$	*DDX5*, whole blood, -0.22, $6.2*10^{-108}$ *CEP95,* whole blood, 0.13, $1.3*10^{-16}$

4.4.5 Comparison to methWAS EWAS results

At cg02598441 (*LOC146880*), the direction of effect estimated in our first (forward) MR analysis was concordant with that in the methWAS EWAS study, with an EWAS P-value ($2.4x10^{-3}$) that survived Bonferroni correction for five tests. The direction of estimated effect was also concordant between studies at the three CpGs at *NTN1*, but the smallest EWAS P-value was 0.12. At cg11398452 (*VAX1*), the direction of estimated effect was discordant between our MR analysis and the methWAS EWAS, with a small EWAS P-value ($9.4x10^{-3}$) (**Table 14**).

4.4.6 Tissue and cleft-subtype-specific variation in the Cleft Collective

At most of the five identified CpGs, there was evidence of weak correlation between methylation in blood, lip and palate tissues (correlation coefficients ranged from -0.11 to 0.32), particularly between blood and lip tissue. Weak evidence was found for the mean methylation values in any of the three tissues differing between the OFC subtypes. However, the analysis was likely underpowered to give precise correlation estimates (**Table 14**).

Table 14: Comparison to methylation data in blood samples from children with an orofacial cleft.

SNP (annotated gene)	**CpG (annotated gene)**	**Forward MR:** *Effect size [1] [standard error]; P-value)*	**methWAS nsCL/P EWAS:** *Effect size [2] [standard error]; P-value)*	**Correlation between blood and lip in the Cleft Collective:** *Correlation coefficient (95% C.I.)*	**Correlation between blood and palate in the Cleft Collective:** *Correlation coefficient (95% C.I.)*	**P-value for difference in mean blood DNA methylation between CLO, CLP and CPO in the Cleft Collective**
rs4752028 (*SHTN1*)	cg11398452 (*VAX1*)	-0.8 [0.1]; $8.7x10^{-9}$	0.01 [0.05]; $9.4x10^{-3}$	0.20 (-0.01, 0.41)	0.07 (-0.00, 0.14)	0.148
rs8076457 (*NTN1*)	cg01862363 (*NTN1*)	-0.8 [0.2]; $3.0x10^{-7}$	-0.030 [0.037]; $2.1x10^{-1}$	-0.04 (-0.24, 0.16)	-0.02 (-0.26, 0.22)	0.828
	cg02481697 (*NTN1*)	-0.6 [0.1]; $3.0x10^{-7}$	-0.023 [0.024]; $1.7x10^{-1}$	0.29 (0.09, 0.49)	-0.06 (-0.35, 0.23)	0.286
	cg16107528 (*NTN1)*	-0.7 [0.1]; $3.0x10^{-7}$	-0.019 [0.016]; $1.2x10^{-1}$	0.34 (0.14, 0.54)	-0.11 (-0.37, 0.15)	0.646
rs1808191 (*CEP95*)	cg02598441 (*LOC146880*)	0.4 [0.1]; $4.3x10^{-7}$	0.020 [0.007]; $2.4x10^{-3}$	0.32 (0.12, 0.52)	0.13 (-0.15, 0.41)	0.548
rs1991401 (*PLEKHM1P1*)						

1 Effect size for forward MR can be interpreted as the difference in risk of nsCL/P per S.D. increase in methylation beta value.
2 Effect size for the methWAS EWAS can be interpreted as the difference in mean methylation beta value in participants with nsCL/P compared to controls.

4.5 Discussion

In this study, a previously devised framework was employed to explore putative mediation of genetic influences on nsCL/P by DNA methylation. Five CpG sites were identified, in three independent regions (*VAX1, LOC146880, NTN1*), where either bidirectional MR and co-localisation analyses suggested that the same variant affects both methylation and nsCL/P, or where evidence was found that two independent variants affect both nsCL/P and methylation (see **Figure 6**).

Lower methylation at the CpG at *VAX1* (cg11398452) was associated with the nsCL/P risk allele C of the SNP rs4752028 at *SHTN1*. This SNP is strongly associated with lower expression of *SHTN1* according to two eQTL databases. *VAX1* is a homeobox containing gene that has been shown to be expressed in the developing brain [279,280] and SNPs in *VAX1* have been shown to be associated with nsCL/P in multiple independent GWAS across distinct populations [13,23,29,281,282]. *VAX1* knock-out mice have been shown to develop a cleft palate (CP), suggesting *VAX1* has a potentially important role in nsCL/P aetiology [13]. *SHTN1***,** sometimes known as *KIAA1598*, codes for the protein (shootin1) that is involved in neuronal polarization [283] and has also been reported to be relevant to the aetiology of nsCL/P in several studies [29,284,285]. It is difficult to discern the locus more relevant to nsCL/P between the *VAX1* and *SHTN1* genes because of their close proximity and similar expression profiles in mice and it is unclear which is the functional gene in the area [29,280,285,286]. The association between methylation at cg11398452 and rs4752028 was not replicated in the GOYA data, but the SNP was strongly associated with methylation at nearby probes. Similarly, the direction of association between cg11398452 methylation and nsCL/P was discordant between the MR analysis and a previously published observational EWAS [36].

Higher methylation at the CpG at *LOC146880* (cg02598441) was associated with the G allele of rs1991401 in *DDX5* and the C allele of rs1808191 in *PLEKHM1P*. rs1991401 in *DDX5* was associated with reduced expression of *DDX5* and increased expression of *CEP95* and *MILR1* while rs1808191 in *PLEKHM1P* was associated with increased expression of *SMURF2* but decreased expression of *PLEKHM1P*, *RP13-104F24.3* and *has-mir-6080*. However, there was weak evidence that the SNPs affected expression of the same genes. *DDX5* is involved in RNA helicase processes that are highly relevant to important cellular processes while *PLEKHM1P* and *LOC146880* are pseudogenes [280]. There is no robust evidence from previous literature to support an association between genetic variation of these genes and nsCL/P. The SNP in *PLEKHM1P* replicated as an mQTL in the GOYA dataset but there was not sufficient data to test the SNP in *DDX5*.

Lower methylation at three CpGs at *NTN1* was found to be associated with the nsCL/P risk allele T of the SNP rs8076457, an intergenic SNP close to *NTN1*. rs8076457, the mQTL for cg08162363, cg02481697 and cg16107528, is not known to be strongly associated with gene expression levels in two datasets. The function of *NTN1* is still largely unknown but is thought to be involved in cell migration during development [280]. *NTN1* has been previously discussed as a strong candidate gene for nsCL/P [30]; *NTN1* may affect liability to nsCL/P via epistatic interactions, there is some evidence that *NTN1* knock-out mice show consistency with the CP phenotype and *NTN1* expression is localised to the palate [30,31]. rs8076457 replicated as an mQTL across all relevant CpGs in the GOYA dataset.

Previous work has identified many functional possibilities for genetic risk variants for nsCL/P [31,32,231] but this study adds to the current evidence for DNA methylation playing a role in the aetiology of nsCL/P. Additional strengths of this

study include the integration of multiple data sources, such as ALSPAC, which provided access to detailed phenotype, genotype and epigenetic data. The nsCL/P GWAS summary statistics allowed a comprehensive genome-wide analysis in a large dataset. The methWAS cohort EWAS results allowed a comparison of the influence of methylation on nsCL/P between observational and MR studies. The use of the GOYA replication cohort allowed triangulation of evidence for mQTL across different studies. Finally, the Cleft Collective data allowed us to compare genome-wide DNA methylation in different tissues and subtypes of cleft.

There are, several limitations to this study. First, methylation and expression in the studied tissues (postnatal cord blood, whole blood, lip and palate tissue) are unlikely to accurately reflect that in the developing orofacial tissue where epigenetic processes could feasibly influence susceptibility to nsCL/P. However, a previous study has identified a high correlation between blood and lip tissue, both taken at the time of first surgery in a UK cohort of patients with non-familial nsCL/P [36]. Previous analysis looking for tissue-specific signals for nsCL/P did not find evidence of enrichment and concluded that this may be due to tissue type differences [254]. Second, CLO and CLP cases were analysed together as one group in the GWAS, MR analyses and the previously published EWAS. Increasingly, evidence suggests that these subtypes are molecularly and aetiologically distinct and should be analysed separately [8,254], but the analysis was limited by the data available from previous studies. Although no evidence was found for differential methylation between subtypes at our five identified CpGs, there may be other loci where methylation mediates genetic influences on more specific cleft subtypes. Third, although efforts were made to select only non-syndromic cases for the Cleft Collective analysis, it is difficult to guarantee that no syndromic cases were included,

and children with syndromes may have very different methylation profiles. Fourth, a major limitation of this study is that some of the steps, particularly the reverse MR, are likely to be statistically underpowered. Fifth, as the majority of mQTL were instrumented by just a single genetic variant, it was not possible to distinguish between mediation and horizontal pleiotropy and therefore proposed mediation is putative. Finally, the comparisons between ALSPAC and other cohorts may be affected by technical differences, tissue differences (cord blood vs whole blood), ancestral differences between cohorts, age of participants (newborns vs children over six years old), a lack of statistical power giving rise to spurious associations or the enrichment of GOYA for overweight and obese mothers, which may introduce selection bias. Indeed, although mQTL were largely concordant between ALSPAC and GOYA, the mQTL and CpG site (in *SHTN1*) found to co-localise with liability to nsCL/P in ALSPAC did not replicate in GOYA. Similarly, the CpG site cg11398452, close to *SHTN1*, was directionally discordant between the analysis in ALSPAC and the results of a methWAS EWAS.

In conclusion, analyses identified three putative loci where DNA methylation may mediate genetic susceptibility to nsCL/P. Future work, determining the function of these genes and the epigenetic modulation of their expression relevant to prenatal orofacial development could provide important aetiological insights. One possibility, warranting further investigation, is that identified DNA methylation differences are related to environmental exposures.

Chapter 5: Investigating the shared genetics of nsCL/P and facial morphology

5.1 Abstract

There is increasing evidence that genetic risk variants for nsCL/P are also associated with normal-range variation in facial morphology. However, previous analyses are mostly limited to candidate SNPs and findings have not been consistently replicated.

In this chapter, PRS were used to test for genetic overlap between nsCL/P and seven biologically relevant facial phenotypes. Where evidence was found of genetic overlap, bidirectional MR was used to test the hypothesis that genetic liability to nsCL/P is causally related to implicated facial phenotypes.

Across 5,804 individuals of European ancestry from two studies, strong evidence was found, using PRS, of genetic overlap between nsCL/P and philtrum width; a 1 S.D. increase in nsCL/P PRS was associated with a 0.10 mm decrease in philtrum width (95% C.I. 0.054, 0.146; P = 0.00002). Follow-up MR analyses supported a causal relationship; genetic variants for nsCL/P homogeneously cause decreased philtrum width. The results support a liability threshold model of inheritance for nsCL/P, related to abnormalities in development of the philtrum.

5.2 Introduction

Facial morphology in the general population has been shown to be highly polygenic; genome-wide significant loci have been found for multiple facial phenotypes across diverse ethnic populations [95,287-290]. In some cases, the genes associated with normal-range variation in facial shape have also been implicated in nsCL/P (e.g. *MAFB*) [95]. Likewise, previous studies using candidate SNPs have found

overlap between nsCL/P risk loci and facial phenotypes in the general population [288,291,292]. For example, the strongest nsCL/P GWAS signal, intergenic variant rs987525 on chromosome 8q24, was found to be associated with more than half of the 48 facial phenotypes studied in a European population [288] while in a Han Chinese population, rs642961 in *IRF6* (a major nsCL/P-associated gene) strongly predicted lip-shape variation in females [292]. Shared genetic aetiology between nsCL/P and normal-range in facial variation is further supported by the observation of sub-clinical differences in facial morphology and the increased incidence of lip defects in unaffected relatives of nsCL/P cases [93,96-98]. However, associations between nsCL/P genetic variants and facial morphology have not been consistently replicated, possibly because of methodological differences in measuring facial phenotypes or population differences between cohorts [287].

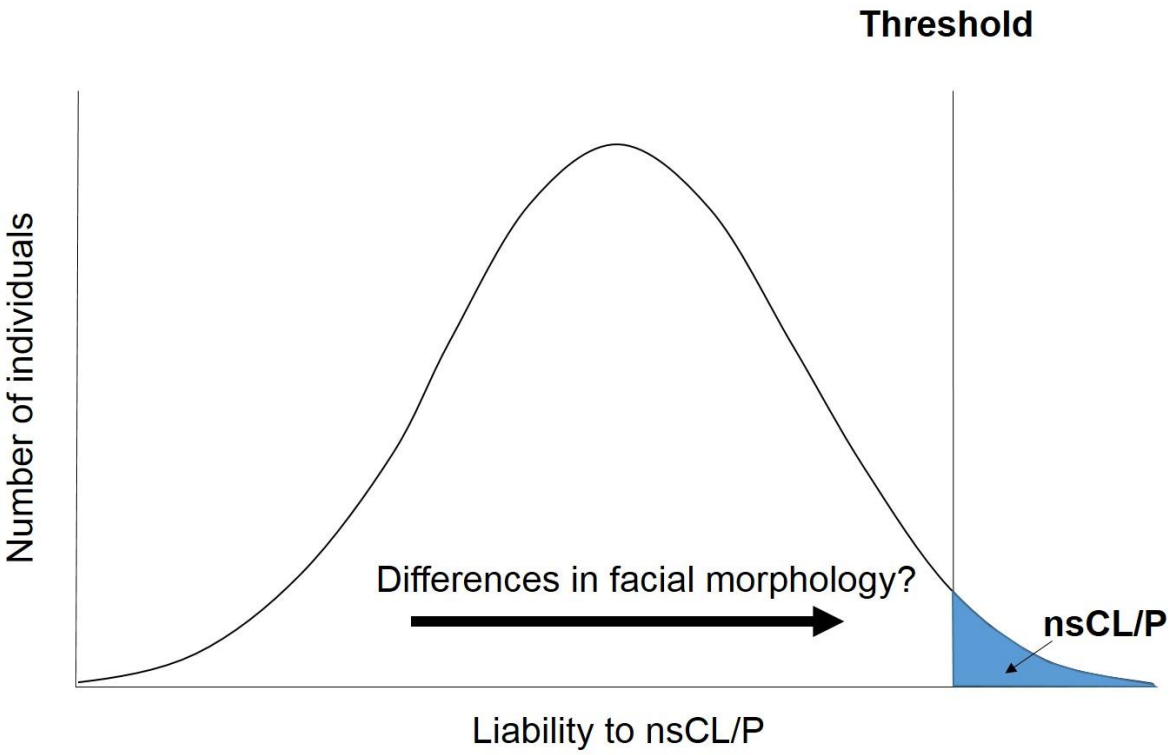
The use of individual markers to demonstrate genetic overlap between two phenotypes has notable limitations; a large number of statistical tests are introduced, and interpretation is difficult when some SNPs show an association and others do not. In **Chapter 3**, evidence was found of over-transmission of nsCL/P PRS between unaffected parents and affected offspring, which suggests that PRS can be used effectively to test for genetic overlap between nsCL/P and normal-range facial morphology.

Interpreting genetic overlap between nsCL/P and a facial phenotype is difficult because the development of the face and development of an OFC are largely synchronous. One possibility is that differences in the facial phenotype are a sub-phenotypic manifestation of genetic liability to nsCL/P (see **Figure 7**). The inheritance of dichotomous traits can be modelled on the liability scale, where every individual has an underlying normally distributed liability to the trait determined by

genes, environment and chance. Individuals above a liability threshold develop the trait, while increased liability may cause related phenotypic differences in individuals without the trait [156,204,293]. A highly relevant example is that increased liability to developing a CP has been hypothesised to be associated with delayed movement of the palatal shelf, which may in turn result in a CP, dependent on other factors such as shelf and head width [204]. Bidirectional MR can be used to test the hypothesis that genetic liability to nsCL/P is causally related to facial morphology [140].

Figure 7: Liability threshold model for nsCL/P

Shown is an illustration of a liability threshold model for nsCL/P. Every individual has a normally distributed liability to nsCL/P, determined by genes, environment and chance. Individuals over the liability threshold develop nsCL/P, with the area under the curve past the threshold equal to the trait incidence. We are hypothesising that liability to nsCL/P, specifically genetic liability to nsCL/P, may be associated with differences in facial morphology across the general population.



The genetic overlap between nsCL/P and normal-range facial morphology in the general population was investigated using nsCL/P PRS. Then, in the instance of genetic overlap, bidirectional MR was used to explore the relationship between nsCL/P and implicated facial phenotypes. In this chapter, Mr Myoung Keun Lee from the University of Pittsburgh performed the polygenic risk scoring analysis in the 3DFN study which required individual level data. Another member of the Pittsburgh group generated the GWAS summary statistics for philtrum width in their cohort. I performed all other described analyses, including the analysis of the 3DFN summary data.

5.3 Methods

5.3.1 Study participants

5.3.1.1 ICC and Bonn-II

The nsCL/P meta-analysis GWAS summary statistics, previously described in **Chapter 3**, were used for information on nsCL/P genetic risk variants. In brief, the meta-analysis summary statistics included ICC TDT results from 638 parent-offspring trios and 178 offspring duos of European descent, meta-analysed with GWAS summary results on 399 cases and 1,318 controls from the Bonn-II study.

5.3.1.2 ALSPAC

Primary analyses used participants from ALSPAC [164]. The cohort and genotyping have previously been described in **Chapter 2**.

Face-shape data are available for a subset of the cohort. ALSPAC children were invited to a clinic at the age of 15 years and 5,253 attended, where two high-resolution 3D facial images were taken by Konica Minolta Vivid 900 laser scanners. 4,747 individuals had usable images (506 individuals did not complete the assessment, or the scans were of poor quality and consequently excluded). The 3D

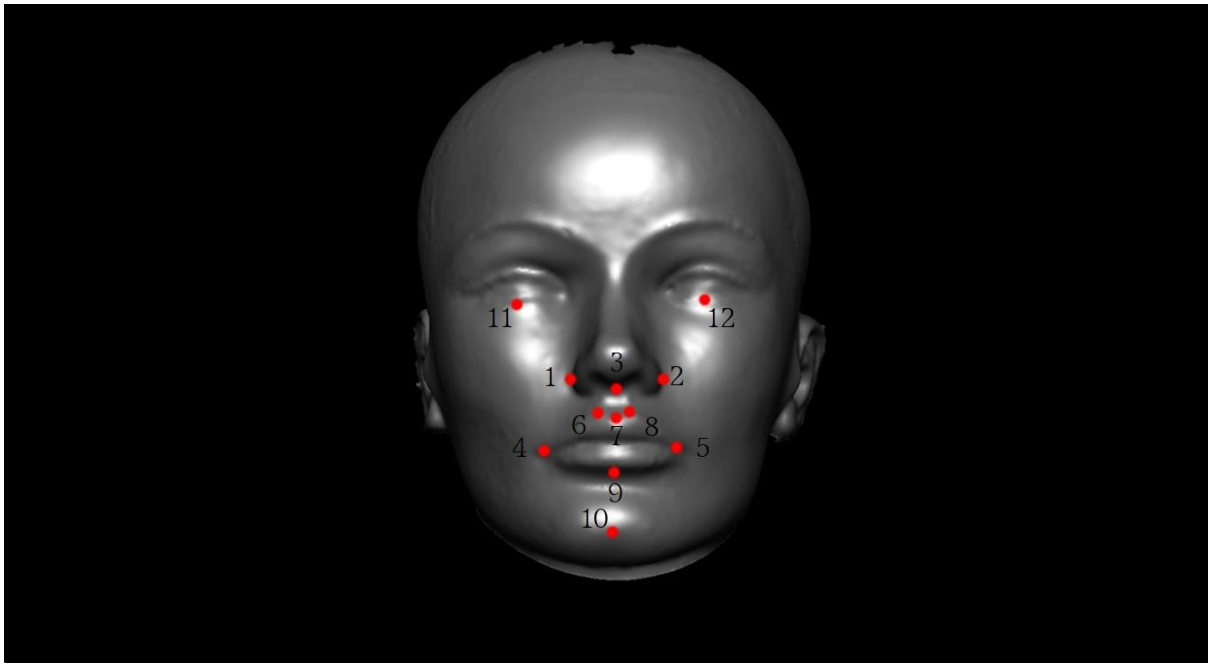
coordinates of 22 facial landmarks were derived from the scans and Euclidean distances between landmarks were calculated for relevant distances. To alleviate multiple testing issues, I tested 7 distances that were either tested previously or have biological relevance to nsCL/P (**Table 15**). Facial distances used in the analysis are shown in **Figure 8**.

Table 15: Biologically plausible facial phenotypes

Facial phenotype	**Justification for inclusion**
Nasal width	Boehringer et al (EJHG 2011) – looked for association between phenotype and nsCL/P SNPs.
Nasal lip height	Relevance to cleft lip
Lip width	Relevance to cleft lip
Philtrum width	Relevance to cleft lip
Lip height	Relevance to cleft lip
Lip chin height	Weinberg et al (Orthodontics & craniofacial research 2009) – compared phenotype between parents of nsCL/P children and control parents.
Inter-palpebral width	Weinberg et al (Orthodontics & craniofacial research 2009) – compared phenotype between parents of nsCL/P children and control parents.

Figure 8: Facial morphological distances of interest

This figure shows the 12 facial landmarks that were used to generate the facial phenotypes tested for association with the nsCL/P PRS. Facial phenotypes were defined as the 3D Euclidean distance between the following landmarks (Nasal width: 1-2, Nasal-lip distance: 3-7, Lip width: 4-5, Philtrum width: 6-8, Lip height: 7-9, Lip-chin distance: 9-10 and inter-palpebral width: 11-12).



5.3.1.3 3D Facial Norms Study (3DFN)

Replication analyses used data on participants from the 3DFN [185] which has been described in **Chapter 2**. In brief, the 3DFN includes 2,454 unrelated individuals of recent European descent, aged between 3 and 40 years, which were recruited from 4 sites across the USA and screened for a history of craniofacial conditions. 3D-derived anthropometric measurements, 3D facial surface images and genotype data were derived from each study participant.

In collaboration with the Center for Inherited Disease Research (CIDR), 2,447 subjects in the 3DFN database were genotyped using a genome-wide association array including 964,193 SNPs from the Illumina OmniExpress+exome v1.2 array and an additional 4,322 SNPs from previous craniofacial genetic studies. The genotype dataset was imputed using the 1000 Genomes reference panel (phase 3) [185].

3DFN study participants had their facial surfaces captured via 3D stereo-photogrammetry using either a two-pod 3dMDface or a multi-pod 3dMDcranial system. Captures were inspected to ensure 3D surface quality and additional captures were obtained if necessary. Similar to ALSPAC, a set of standard facial landmarks were collected from each 3D facial image and linear distances were calculated between the landmark coordinates.

5.3.2 Polygenic risk score construction and analysis

5.3.2.1 PRS construction

The nsCL/P meta-analysis GWAS summary statistics were used to construct nsCL/P PRS. The PTDT results from **Chapter 3** were used to inform the most predictive PRS. The P-value inclusion threshold, which was P<0.00001, was selected based on the most predictive threshold in the European trios.

Using ALSPAC as a reference panel for linkage disequilibrium, PLINK was used to prune and clump the nsCL/P meta-analysis summary statistics (r^2<0.1 and 250 kb) using the most predictive P-value threshold. The PRS were then constructed in the ALSPAC sample. Information on the SNPs in the PRS is contained in **Table 16**.

Table 16: nsCL/P Polygenic risk score SNPs

SNP	CHR:BP[1]	Effect Allele	Other Allele	nsCL/P Beta	nsCL/P S.E.	nsCL/P P-value
rs742071	1:18979874	T	G	0.2927	0.0573	3.3×10^{-7}
rs560426	1:94553438	T	C	-0.2574	0.0564	5.1×10^{-6}
rs4147812	1:94575043	A	C	0.3234	0.0606	9.4×10^{-8}
rs12057415	1:94829769	T	C	-0.2786	0.0572	1.1×10^{-6}
rs861020	1:209977111	A	G	0.3218	0.0660	1.1×10^{-6}
rs7590268	2:43540125	T	G	-0.3428	0.0650	1.3×10^{-7}
rs1650504	5:158029550	A	G	0.2585	0.0584	9.5×10^{-6}
rs12543318	8:88868340	A	C	-0.288	0.0594	1.3×10^{-6}
rs6470648	8:129716308	A	G	0.3081	0.0636	1.3×10^{-6}
rs11989880	8:129872982	T	C	0.4993	0.0586	1.5×10^{-17}
rs12548036	8:129947882	T	G	0.5416	0.0585	2.1×10^{-20}
rs1372452	8:130029034	A	G	0.5244	0.0778	1.6×10^{-11}
rs3138512	9:92222453	A	G	0.3085	0.0690	7.9×10^{-6}
rs4752028	10:118834991	T	C	-0.4046	0.0703	8.8×10^{-9}
rs9545330	13:80699166	A	G	0.432	0.0684	2.6×10^{-10}
rs1258763	15:33050423	T	C	0.3049	0.0629	1.2×10^{-6}
rs1873147	15:63312632	A	G	-0.3518	0.0621	1.4×10^{-8}

rs8076457	17:8943929	T	C	0.3175	0.0620	3.1×10^{-7}
rs227731	17:54773238	T	G	-0.3148	0.0564	2.5×10^{-8}
rs1808191	17:62784028	A	C	-0.3307	0.0707	2.9×10^{-6}
rs3746101	19:2050823	T	G	0.4474	0.0980	5.0×10^{-6}

1 CHR:BP – Chromosome and Base Pair Position on the HumanGenome19 build

5.3.3.2 PRS power calculations

Power calculations for PRS analysis were performed using AVENGEME [136,224]. Assuming 80% power and an alpha level of 0.05, we estimated the minimum genetic covariance required between nsCL/P and the 3D face-shape distances for an association between the PRS and the face-shape distances to be detectable. Parameters used in power calculations are contained in **Table 17**. The genetic covariance estimates were then converted to genetic correlation estimates using GCTA [220] heritability estimates of the facial morphology variables derived in ALSPAC.

Table 17: Parameters in Polygenic Risk Score analysis power calculations

Parameter	**Value**	**Source**
Sample size of training sample	3972	determined from data
Prevalence of nsCL/P in training sample	0.305	determined from data
Population prevalence of nsCL/P	0.001	IPDTOC Working Group. "Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC)." (2011).
Number of independent SNPs common to both arrays (r^2<0.1)	75,737	determined from data
h^2 of nsCL/P	0.2	AVENGEME estimate
Proportion of null markers	0.992	AVENGEME estimate

5.3.3.2 Polygenic risk score association with facial phenotypes in ALSPAC

Of the 4,747 ALSPAC children with face-shape scans, 3,941 individuals had genotype data. GCTA [220] was used to prune these individuals for relatedness (GRM < 0.05) and the final sample with complete covariates included 3,707 individuals. The associations between facial phenotypes and the nsCL/P PRS were measured in the final sample using a linear regression adjusted for sex, age at clinic visit, height at clinic visit and the first four principal components. Effect sizes were reported per S.D. increase in PRS.

5.3.3.3 Replication in 3D Facial Norms Database

Distances with some evidence of an association ($P < 0.05$) in the ALSPAC children were followed up for replication in an independent cohort (3DFN). 2,429 3DFN individuals had genotype and face-shape data. 332 individuals were removed due to missing SNPs in the PRS or missing covariates. The final sample consisted of 2,097 individuals. The association between implicated facial measurements and the nsCL/P PRS was measured using a linear regression adjusted for sex, age, height and the first 4 principal components. Effect sizes were reported per S.D. increase in PRS.

5.3.4 Exploring possible mechanistic direction

5.3.4.1 Bidirectional Mendelian randomization analysis

A bidirectional two-sample MR analysis was performed using the TwoSampleMR R package [273] , testing both the forward direction (the effect of genetic risk variants for nsCL/P on implicated facial measurements) and the reverse direction (the effect of genetic risk variants for implicated facial measurements on liability to nsCL/P).

The IVW method was used as the primary analysis (previously described in **Chapter 4**). Several sensitivity analyses were performed to test the assumptions of MR. Cochran's Q test for heterogeneity was used to test for balanced pleiotropy by testing the heterogeneity of the Wald estimates across each SNP [294,295]. MR Egger was used to test for directional pleiotropy; the MR Egger regression estimate provides an estimate for the true causal effect even if all genetic variants are invalid but assumes that instrument strength and pleiotropic effects are independent (the INSIDE assumption).

The weighted median [148] and weighted mode estimates [149] were used to further examine the consistency of the causal estimate under certain assumptions. Assuming that at least half of the genetic instruments are valid, the weighted median provides a causal estimate, allowing for violation of the second and third IV assumptions. An advantage of the weighted median is that it generates more precise causal estimates than MR Egger [148]. Assuming that the modal estimate (the largest grouping of similar causal effect estimates) is valid, the weighted mode method estimates the causal effect even if all other instruments are invalid. The weighted mode is generally less precise than the weighted median but more precise than MR-Egger [149].

Leave-one-out analysis, where the IVW analysis is rerun after removing each SNP in turn, was used to identify potential outlying SNPs that are driving the overall effect estimate. The Steiger test of directionality [296] was used to determine the likely direction of effect using GWAS summary statistics data. The test is particularly useful, compared to a bidirectional MR, when the biology of instruments is unknown or one of the exposure or outcome has no valid instruments

5.3.4.2 GWAS summary statistics for nsCL/P and implicated facial phenotypes

MR analyses required relevant SNP association information with respect to both nsCL/P and implicated facial measurements. SNP information relevant to nsCL/P was extracted from the nsCL/P meta-analysis summary statistics, previously described in **Chapter 3**.

For implicated facial phenotypes, GWAS were performed using PLINK [161] in both ALSPAC (3,707 individuals) and the 3DFN study (2,429 individuals with genotype and face shape data), using the same covariates as previously described in the PRS analysis**.** Summary statistics were then meta-analysed using METAL [241], with the combined sample including 6,136 individuals. SNP information relevant to implicated facial phenotypes was then extracted from the ALSPAC/3DFN meta-analysis summary statistics.

The ALSPAC/3DFN meta-analysis GWAS summary statistics of implicated facial phenotypes were subsequently analysed and functionally annotated [297] with the potential overlap between implicated facial phenotype associated SNPs and eQTL investigated using the GTEx catalogue, described previously in **Chapter 2** [183].

5.3.4.3 Genetic risk variants for nsCL/P and implicated facial phenotypes

For the forward direction, relevant SNPs are variants strongly associated with nsCL/P. Six of twelve genome-wide significant SNPs from a previous study were used as nsCL/P SNPs [24]. The study was a meta-analysis of both European and Asian data and these six SNPs were genome-wide significant in the European only meta-analysis. Information on the nsCL/P SNPs is contained in **Table 18**.

Table 18: nsCL/P Mendelian randomization SNPs

SNP	CHR:BP[1]	Effect Allele / Other Allele	nsCL/P Beta	nsCL/P S.E.	Philtrum width Beta	Philtrum width S.E.
rs7590268	2:43540125	T/G	-0.328	0.065	0.0537	0.0505
rs987525	8:129946154	A/C	0.829	0.0909	-0.1045	0.0517
rs7078160	10:118827560	A/G	0.400	0.0704	0.0806	0.0572
rs8001641	13:80692811	A/G	0.357	0.0581	-0.0021	0.0432
rs1873147	15:63312632	A/G	-0.352	0.0621	0.0491	0.0483
rs227731	17:54773238	T/G	-0.315	0.0564	0.0929	0.0434

1 CHR:BP – Chromosome and Base Pair Position on HG19

For the reverse direction, relevant SNPs are variants strongly associated with the implicated facial phenotypes. The ALSPAC/3DFN meta-analysis summary statistics were LD clumped (r2<0.001 within 500KB) to generate independent instruments for the MR analysis. LD proxies (r2>0.9) were used for SNPs unavailable in the nsCL/P summary statistics and were generated using LDlink and LDproxy [277] using the 1000 Genomes CEU/GBR populations as the reference panel [163].

5.3.4.4 Interpretation of bidirectional MR analysis

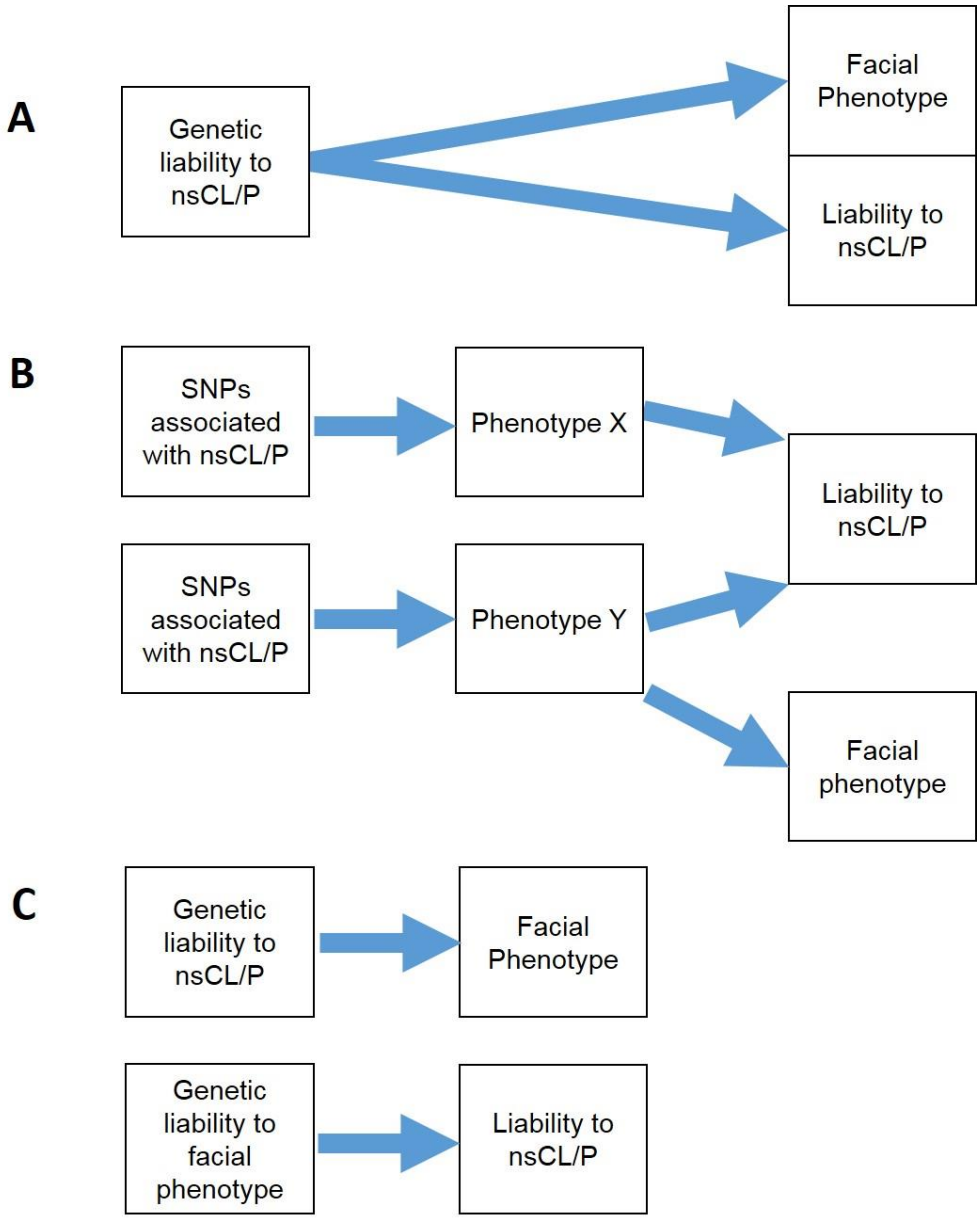
The results of the bidirectional MR and relevant sensitivity analyses were used to infer the likelihood of the liability-related sub-phenotype model. Three distinct possibilities were considered to explain the association between nsCL/P PRS and implicated facial phenotypes (see **Figure 9**).

Figure 9: Interpretation of bidirectional MR

(A) *SNPs associated with nsCL/P have a homogeneous effect on the facial phenotype with weak evidence for the reverse direction MR*. We would conclude that genetic liability to nsCL/P causes both increased liability to nsCL/P (in conjunction with the environment and chance) and changes in the facial phenotype.

(B) *SNPs associated with nsCL/P have a heterogeneous effect on the facial phenotype*. In this instance, there is weak evidence for genetic liability to nsCL/P causing changes in the facial phenotype because liability assumes a consistent effect. We would conclude that an unknown confounder Y affects the facial phenotype and liability to nsCL/P independently.

(C) *SNPs associated with nsCL/P have a homogeneous effect on the facial phenotype AND SNPs associated with the facial phenotype cause increased liability to nsCL/P*. In this instance, there are two possibilities. The first possibility is that the genetic instruments for the facial phenotype are weak (e.g. only one SNP) and so the causal effect estimate of the facial phenotype on liability to nsCL/P is imprecise. The second possibility is that nsCL/P and the facial phenotype have a substantial genetic correlation, which would require further investigation. Here, the results of the Steiger test are useful, as they can infer the most likely direction of effect between nsCL/P and implicated facial phenotypes.



5.4 Results

5.4.1 The prediction of facial morphology using PRS for nsCL/P

Prior to testing the performance of our nsCL/P PRS on predicting facial morphology, the minimum genetic correlations required to detect an association between the PRS and the facial phenotypes were calculated using AVENGEME. Across the facial phenotypes, the minimum genetic correlation required ranged from 0.17 to 0.28 with differences attributable to the different heritability estimates across the facial phenotypes (**Table 19**).

Table 19: Power calculations for polygenic risk scoring

3D facial Euclidean distances in ALSPAC (N=3737)	**AVENGEME: Minimum Genetic Covariance**	**SNP heritability (h^2) estimate from GCTA (95% C.I.)**	**Minimum Genetic Correlation between nsCL/P and developmental outcome detectable[1] (95% C.I. from h^2 estimates)**
Lip width chL_chR	0.045	0.18 (0.04, 0.33)	0.24 (0.18, 0.50)
Philtrum width cphL_cphR	0.045	0.20 (0.05, 0.35)	0.23 (0.17, 0.45)
Lip height ls_li	0.045	0.18 (0.03, 0.32)	0.24 (0.18, 0.58)
Nasal width alL_alR	0.045	0.13 [0.00, 0.27)	0.28 (0.19, 1]
Distance between eyes piL_piR	0.045	0.34 (0.19, 0.49)	0.17 (0.14, 0.23)
Distance between lip and chin li_pg	0.045	0.21 (0.06, 0.35)	0.22 (0.17, 0.41)
Distance between lip and nose sn_ls	0.045	0.34 (0.19, 0.49)	0.17 (0.14, 0.23)

[1] $Genetic\ Correlation_{AB} = \frac{Genetic\ Covariance_{AB}}{\sqrt{h_A^2 h_B^2}}$

The performance of the nsCL/P PRS for prediction of seven facial morphological traits was evaluated. Evidence was found of an association between the nsCL/P PRS and philtrum width in the ALSPAC children, where a 1 S.D. increase in nsCL/P PRS was associated with a 0.07 mm decrease in philtrum width (95% C.I. 0.02, 0.13; P=0.014) (**Table 20**).

Table 20: Association of nsCL/P PRS with facial phenotypes in ALSPAC children

3D facial Euclidean distances in ALSPAC	ALSPAC children (N=3707)	
	Beta (95% C.I.) Per 1 S.D. increase in PRS	P-value
Distance between subnasale and labiale superius (Nasal-lip)	-0.25 (-2.16, 1.65)	0.79
Distance between labiale inferius and pogonion (Lip-chin)	-0.02 (-0.10, 0.06)	0.64
Distance between left and right palpebrale inferius (Mid-point of eyes)	-0.08 (-0.17, 0.01)	0.09
Distance between left and right alare (Nasal width)	-0.01 (-0.08, 0.06)	0.75
Distance between labiales inferius and superius (lip height)	0.02 (-0.05, 0.10)	0.53
Distance between left and right crista philtri (philtrum width)	-0.07 (-0.13, -0.02)	0.014
Distance between left and right cheilion (lip width)	-0.02 (-0.15, 0.10)	0.70

Replication for this finding was attempted in the 3DFN study and a consistent effect of 1 S.D. increase in nsCL/P PRS being associated with a 0.14 mm decrease in philtrum width (95% C.I. 0.07, 0.21; P = 0.00017) was found. Meta-analysing these results; indicated that a 1 S.D. increase in nsCL/P PRS is associated with a 0.10 mm decrease in philtrum width (95% C.I. 0.054, 0.146; P = 0.00002).

5.4.2 GWAS of philtrum width

To generate SNP-philtrum width association information for MR analyses, GWAS of philtrum width in both ALSPAC and 3DFN were performed separately before meta-analysing. The combined sample included 6,136 individuals of recent European descent. Two novel chromosomal regions associated with philtrum width with genome-wide significance were identified at 5q22.2 (lowest P value for rs255877, $P=3.8x10^{-10}$), within the non-coding RNA intronic region of an uncategorised gene *ENSG00000232633*, and 7p15.2 (rs2522825, $P=1.4x10^{-8}$), an intergenic SNP near *HOXA1*. There was evidence of possible heterogeneity between ALSPAC and 3DFN with regards to the association between rs2522825 and philtrum width. Although the direction of effect was consistent between the two studies, confidence intervals did not overlap (**Table 21**).

Table 21: Independent philtrum width trait loci derived from the ALSPAC/3DFN summary statistics

Variant	CHR:BP	Effect allele / Other allele	ALSPAC (N=3707)		3DFN (N=2429)		Combined meta-analysis (N=6136)	
			Beta (95% C.I.)	P-value	Beta (95% C.I.)	P-value	Beta (95% C.I.)	P-value
rs255877	5:112753584	G/A	0.24 (0.15, 0.32)	$6.2x10^{-8}$	0.16 (0.07, 0.25)	$8.1x10^{-4}$	0.20 (0.14, 0.26)	$3.8x10^{-10}$
rs2522825	7:27111994	T/C	-0.30 (-0.39, -0.20)	$2.8x10^{-10}$	-0.08 (-0.17, 0.02)	0.11	-0.19 (-0.26, -0.13)	$1.4x10^{-8}$

Some evidence was found that the two lead SNPs may be eQTL for nearby genes (**Table 22**). The two lead SNPs of the genome-wide significant loci, rs255877 and rs2522825, were used as genetic variants associated with philtrum width in subsequent MR analyses.

Table 22: Philtrum width associated SNPs in GTex

SNP (Effect allele)	Gene	Effect size of association with gene expression (P-value)	Tissue
rs255877 (G)	*YTHDC2*	-0.32 ($3.4x10^{-9}$)	Brain – Cerebellum
		-0.11 ($2.1x10^{-6}$)	Thyroid
		-0.14 ($2.4x10^{-5}$)	Lung
	CTD-2201G3.1	0.23 ($3.6x10^{-5}$)	Skin – sun exposed (lower leg)
rs2522825 (T)	*SKAP2*	0.20 ($3.2x10^{-14}$)	Whole blood
	HOXA4	0.43 ($2.2x10^{-9}$)	Whole blood
	HOXA-AS2	0.41 ($7.0x10^{-9}$)	Whole blood
	HOTAIRM1	0.35 ($3.1x10^{-8}$)	Whole blood
	HOXA2	0.37 ($4.4x10^{-8}$)	Whole blood
	HOXA5	0.31 ($4.0x10^{-6}$)	Skin – not sun exposed (suprapubic)
	HOXA1	0.33 ($1.2x10^{-5}$)	Whole blood
		0.24 ($2.1x10^{-5}$)	Skin – sun exposed (lower leg)
	HOXA6	0.24 ($4.3x10^{-5}$)	Skin – not sun exposed (suprapubic)

5.4.3 Bidirectional MR

MR was used to investigate the possible causal mechanism that would give rise to the genetic overlap between nsCL/P and philtrum width.

Firstly, it was determined whether genetic variants contributing to liability of nsCL/P affect philtrum width, by testing SNPs strongly associated with nsCL/P for association with philtrum width. A 1-unit log odd increase in liability to nsCL/P was associated with a 0.11mm (95% C.I. 0.04, 0.19; P = 0.0036) decrease in philtrum

width. Sensitivity analyses suggested there was weak evidence for pleiotropy or heterogeneity and validated the consistency of the instrument. Leave-one-SNP-out analysis showed consistent effect estimates after exclusion of each SNP (**Table 23**).

Table 23: Causal estimates of genetic liability for nsCL/P on philtrum width using Mendelian Randomization and sensitivity analyses.

Test	**Interpretation**	**Estimate (95% C.I.)**	**P-value**
Inverse variance weighted	Primary causal estimate[1]	-0.11 (-0.19, -0.04)	0.0036
Heterogeneity of Inverse variance weighted	Balanced pleiotropy	N/A	0.36
MR-Egger	Intercept test for directional pleiotropy[2]	-0.01 (-0.11, 0.10)	0.93
	Regression estimate[1]	-0.10 (-0.33, 0.13)	0.43
Weighted median	Consistency[1]	-0.12 (-0.21, -0.04)	0.0043
Weighted mode	Consistency[1]	-0.12 (-0.21, -0.03)	0.049
Leave-one out rs1873147	Additive model[1]	-0.11 (-0.20, -0.02)	0.017
Leave-one out rs227731	Additive model[1]	-0.10 (-0.16, -0.03)	0.007
Leave-one out rs7078160	Additive model[1]	-0.13 (-0.20, -0.06)	0.0001
Leave-one out rs7590268	Additive model[1]	-0.11 (-0.20, -0.02)	0.013
Leave-one out rs8001641	Additive model[1]	-0.13 (-0.21, -0.04)	0.0030
Leave-one out rs987525	Additive model[1]	-0.10 (-0.22, 0.01)	0.084

1 Units: mm change in philtrum width per 1-unit log odd increase in liability to nsCL/P

2 Units: Average pleiotropic effect of a nsCL/P genetic variant on philtrum width

Secondly, it was determined whether genetic variants associated with philtrum width also affect liability to nsCL/P, by testing two independent SNPs associated with philtrum width at genome-wide significance (derived in the ALSPAC and 3DFN cohorts) for association with nsCL/P**.** Utilising strong LD proxies (**Table 24**), weak

evidence was found of an association between philtrum width-associated variants and liability to nsCL/P. A 1 mm increase in philtrum width was associated with a 0.30 log-odd unit increase in nsCL/P (95% C.I. -0.26, 0.86; P = 0.30). The direction of effect was discordant to the PRS analysis and the MR analysis with liability to nsCL/P as the exposure, where the results suggested that increased liability to nsCL/P is associated with decreased philtrum width. Sensitivity analyses for pleiotropy were limited, with only 2 SNPs.

Table 24: Proxy SNPs (for philtrum width associated variants) in nsCL/P summary statistics

SNP	**Proxy SNP / 1000G CEU & GBR r^2**	**Proxy CHR:BP**	**Effect allele / Other allele**	**Philtrum width Beta**	**Philtrum width S.E.**	**nsCL/P Beta**	**nsCL/P S.E.**
rs255877	rs13188946 / 0.97	5:112722855	T/C	0.20	0.032	0.0088	0.058
rs2522825	rs2712248 / 0.95	7:27120689	T/C	-0.18	0.039	-0.11	0.061

Thirdly, the MR-Steiger test of directionality was used to test the direction of effect between philtrum width and liability to nsCL/P. The results suggested that the true direction of effect is that genetic variants contributing to liability to nsCL/P cause changes in philtrum width ($P < 10^{-10}$).

5.4.4 Interpretation of Bidirectional Mendelian randomization

Strong evidence was found for genetic liability to nsCL/P causing decreased philtrum width, weak evidence was found for heterogeneity or assumption violations

in the forward-MR, and weak evidence was found for the reverse-MR of philtrum width-associated variants on liability to nsCL/P. Therefore, the most likely explanation for the genetic overlap between nsCL/P and philtrum width is that genetic liability to nsCL/P is causally related to decreased philtrum width.

5.5 Discussion

In this chapter, strong evidence has been found for genetic overlap between nsCL/P and normal-range variation in philtrum width. Furthermore, genetic risk SNPs for nsCL/P have been shown to consistently cause decreased philtrum width in the general population. Notably there was weak evidence for genetic overlap between nsCL/P and upper lip width despite the observational correlation between the widths of the upper lip and philtrum.

There are two main implications of these results. First, the findings demonstrate the aetiological relevance of the formation of the philtrum to nsCL/P. The medial nasal and maxillary processes are responsible for development of the upper lip and philtrum [2]. Developmental anomalies within these processes may result in a cleft lip (CL) [3] and the findings show that even when there is successful fusion, as in our study populations, the genetic variants which give rise to a CL/P cause decreased philtrum width. Secondly, the additive effect of common nsCL/P risk variants, on a related phenotype in the general population, showing little evidence for effect heterogeneity, supports a polygenic threshold model of inheritance for nsCL/P.

Although previous studies have looked at nsCL/P related sub-phenotypes, the analyses in this chapter use causal inference methods to more formally investigate

the relationship. The identification of phenotypic differences related to nsCL/P liability are consistent with previous studies [93,96-98,102,298] observing sub-clinical facial phenotypes in individuals with nsCL/P and their unaffected family members, particularly a previous study which observed reduced philtrum width in unaffected parents of individuals with nsCL/P [98]. A polygenic threshold model of inheritance related to development of the philtrum is consistent with a previously proposed mechanism for the inheritance of CP [204], the identification of numerous common nsCL/P genetic risk variants [24-28] and estimation of a substantial SNP heritability for nsCL/P [26]. Previously reported associations between nsCL/P and other facial morphological dimensions found in previous studies [98,288,291] using candidate SNPs were not replicated. However, polygenic risk score methods are methodologically distinct and are used to investigate a different research question to single SNP analyses.

The investigation of the association between nsCL/P and facial morphology was extended in two important ways. The association was shown to be present not only in unaffected family members but also in the general population, and MR was used to demonstrate that this relationship is present on the liability scale. Conventionally MR is used to test possible causal effects of a modifiable continuous exposure such as cholesterol or alcohol on disease outcomes [154,155]. Here the principles of MR were exploited to test the threshold hypothesis, by inferring a causal relationship between genetic variants contributing to liability of nsCL/P and philtrum width in a non-clinical population. The evidence of a causal relationship suggests that a smaller philtrum width is a sub-phenotypic manifestation attributable to the same genetic variants that cause nsCL/P.

In addition to investigating the relationship between facial morphology and nsCL/P, I also performed the first GWAS of philtrum width, and identified two novel genome-wide significant loci. Notably one of the loci, rs2522825 at 7p15.2, was associated with gene expression at several nearby genes in the homeobox gene family, which are known to play important roles in embryonic development [299,300].

The causal inference made in this chapter was achieved through the use of two independent cohorts as discovery and replication samples which greatly reduces the risk of false positives and demonstrates that results can be generalised to different populations. Detailed facial phenotyping data on a large number of individuals in our cohorts along with other detailed phenotype and genotype data enabled us to identify philtrum width as being the most relevant facial morphological feature from amongst seven biologically likely candidates. Statistical power does limit the detection of other features that may have mechanistic relationships with smaller effect sizes.

In this study, I combined CLP and CLO, however there is evidence suggesting that there are distinct aetiological differences between these traits, [8,9,27] which could reduce our statistical power, and complicates interpretation. For example, the philtrum may be more related to CLO, but I did not have sufficient data to compare nsCL/P subtype differences. An additional limitation is that there are few well-characterised genetic risk loci for philtrum width, so the MR analysis testing if genetic variants associated with a narrow philtrum width also affect liability of nsCL/P, may be underpowered. Additionally, one of the two novel identified genome-wide significant SNPs for philtrum width in the meta-analysis had non-overlapping confidence intervals between the two data-sets. This suggests heterogeneity and the possibility that the SNP may be a weak genetic instrument for philtrum width.

The conclusion of this chapter is that genetic liability to nsCL/P is causally related to variation in philtrum width. This finding supports a polygenic threshold model of inheritance for nsCL/P, related to abnormalities in development of the philtrum. Further research looking at the relationship between genetic liability for nsCL/P and severity of cleft would provide further evidence for the polygenic threshold model.

Chapter 6: nsCL/P and adverse developmental outcomes

6.1 Abstract

Syndromic OFC cases often present with other developmental abnormalities in addition to the facial cleft structure while nsCL/P cases do not commonly present with other clinical symptoms. However, there is evidence to suggest that nsCL/P cases may be at increased risk for adverse outcomes such as low birth-weight, dental anomalies, behavioural outcomes, speech disorders and middle-ear disease. The increased incidence of adverse outcomes may only be in OFC cases or it may be present in individuals with a high number of nsCL/P genetic risk variants, for example; in unaffected family members. If genetic variants for nsCL/P are also associated with an adverse outcome, one possibility is that the underlying liability to nsCL/P causes changes in the outcome.

In this chapter, nsCL/P PRS were used to test for genetic overlap between nsCL/P and speech, hearing and anthropometric-related outcomes but there was insufficient power to test for an association with dental variables. Across all phenotypes tested, weak evidence was found of genetic overlap between nsCL/P and speech, anthropometric measures and hearing, although analyses may have been underpowered to detect modest genetic correlation. The results are consistent with the adverse outcomes being caused by an OFC or related surgical procedures rather than being related to genetic liability.

6.2 Introduction

OFC cases caused by a Mendelian syndrome can often present with other developmental or physical abnormalities. Although these abnormalities are not

common in nsCL/P cases, there is substantial evidence that individuals with nsCL/P may be at increased risk for a range of adverse developmental outcomes relating to anthropometrics, tooth formation, hearing, behavioural outcomes and speech. The increased risk of these adverse outcomes means that individuals with nsCL/P may have a reduced quality of life [13,91] and more alarmingly, individuals with nsCL/P have been shown to have higher mortality up to the age of fifty-five [92].

Determining the aetiological factors resulting in increased risk of these adverse outcomes could have important implications for treatment, interventions and biological understanding. I start by reviewing the epidemiological literature, discuss possible causes of the adverse outcomes and propose the use of genetics for disentangling the aetiology.

6.2.1 Anthropometric and dental outcomes

There is some evidence suggesting that individuals with OFCs may differ in terms of physically observable phenotypes (independent from the OFC) compared to the general population. Phenotypic differences include reduced weight, height and head circumference at birth [99-101], and higher incidence of dental anomalies such as hypodontia or a crossbite [102-109]. Over time, the anthropometric differences are thought to attenuate, with some evidence for catch-up growth [99,101]. The same genes may affect dental anomalies and OFCs independently [301]; e.g., a nonsense mutation in *MSX1* has been previously shown to be associated with both OFCs and tooth agenesis in a Dutch family [302].

6.2.2 Speech and hearing

There is also evidence that individuals with an OFC may have increased risk of adverse speech and hearing-related outcomes. Speech, the physical production of

sounds, is an important outcome in OFC patients because an unoperated CP can result in severely disordered speech, with problems also potentially persisting even if the CP is surgically corrected [115]. Common speech disorders affecting OFC children include abnormal nasality, low speech quality and abnormal consonant production [114-116]. Eustachian tube dysfunction is common in children with a CP and can lead to increased risk of middle ear disease and impaired hearing. In particular, otitis media with effusion (OME) is more common with increased severity and a longer duration [110-113].

6.2.3 Behavioural outcomes and education attainment

Children with OFCs may also be more likely to exhibit behavioural problems, such as anxiety, depression and hyperactivity [117-120] and perform worse academically than their peers [121]. Possible reasons for these differences include; difficulties with speech, hearing and facial differences; increased incidence of being bullied and increased absence from school for treatment. These factors may lead to lower self-esteem and impaired emotional development [117].

6.2.4 Possible causes of adverse developmental outcomes

One possible cause of the increased risk of adverse outcomes in nsCL/P cases is the binary (OFC or control) phenotype itself. The physical presence of an OFC could directly affect outcomes, such as the presence of a CP affecting sound production [114-116]. Alternatively, the related surgical procedures that individuals with an OFC may undergo (e.g. palatoplasty and pharyngeal flap surgery) could have adverse effects. Surgery may affect maxillofacial growth [303] which could have a downstream effect on hearing and speech.

Another possibility is that the increased risk of adverse developmental outcome is related to an individual's underlying liability to nsCL/P, independent of whether they have an OFC. Although nsCL/P is treated as a binary phenotype, sub-phenotypes may exist in individuals without an OFC, such as unaffected family members. The possibility that liability to nsCL/P may directly cause the increased incidence of these outcomes, can be explored by testing the association between nsCL/P genetics variants and the outcome in individuals that do not have an OFC. If nsCL/P genetic variants are associated with adverse outcomes in the general population this supports the notion that the outcome is directly related to a continuous nsCL/P phenotype. Conversely, if nsCL/P genetic variants are not associated with tested outcomes, this suggests that previously discussed factors such as the physical presence of an OFC or related surgery may be more aetiologically relevant to these specific outcomes. Determining the most likely aetiological factors for adverse outcomes may have implications for both surgery and the follow-up of patients post-surgery. A similar approach has been previously demonstrated in **Chapter 5**, where strong evidence was found suggesting that increased genetic liability to nsCL/P is associated with decreased philtrum width in the general population.

6.2.5 Chapter aims

In this chapter, I decided to focus on potential genetic overlap between nsCL/P and speech, hearing, dental and anthropometric-related outcomes in analyses. Given the vastly different sample sizes for the different outcomes, power calculations were initially performed to determine which outcomes were feasible to investigate with available data-sets. The genetic overlap between nsCL/P and relevant outcomes was then tested in ALSPAC, using nsCL/P PRS. In the instance

of evidence of genetic overlap, bidirectional MR was used to explore the relationship between nsCL/P and implicated outcomes. In this chapter, I performed all described analyses.

6.3 Methods

6.3.1 Study participants

6.3.1.1 ICC and Bonn-II

The nsCL/P meta-analysis GWAS summary statistics, previously described in **Chapter 3**, were used for information on nsCL/P genetic risk variants. In brief, the meta-analysis summary statistics included ICC TDT results from 638 parent-offspring trios and 178 offspring duos of European descent, meta-analysed with GWAS summary results on 399 cases and 1,318 controls from the Bonn-II study.

6.3.1.2 ALSPAC

Analyses in this chapter used ALSPAC [164] participants with genotype data and phenotype data relevant to anthropometric measures, dental outcomes, speech and hearing. The cohort and genotyping have previously been described in **Chapter 2**.

Anthropometric measures (e.g. birthweight and head circumference) were obtained from 8,677 new-borns registered in the cohort by trained ALSPAC staff. Dental outcomes were measured in ALSPAC children at the Children in Focus clinics. A small sub-sample of ALSPAC were invited to attend specialist clinics at 3 different time-points (31 months, 43 months and 61 months). The earliest time-point of 31 months was used as it had the largest sample size for the dental phenotypes. A binary variable was derived for crossbite status. If an individual had a left or right posterior crossbite they were classified as having a crossbite. If an individual had neither a left nor right crossbite they were classified as a control. Similarly, a binary

variable was derived for missing lateral incisor status. If an individual was missing the upper right lateral incisor and/or the upper left incisor they were classified as having a missing lateral incisor. If both teeth were non-missing they were classified as a control.

ALSPAC children were invited to a clinic at the age of 8 years and 7,390 attended and completed a speech and language assessment. Speech samples were recorded digitally during an expressive language task and listener judgement on the extracted phonetic speech, transcription and comparison with controls was used to diagnose persistent speech disorder (PSD). Participant children were divided into a PSD group (n=263), a group demonstrating common clinical distortions (e.g. problems with r's and s's), a sub-clinical PSD group (n=582), children who presented with a range of speech errors but did not reach the threshold for PSD (n=141) and the rest of the cohort (n=6,399) [304-306]. For the purposes of analyses, two binary variables were constructed from the classifications. Firstly, controls (N=6399) were compared to children diagnosed with PSD (N=263). However, the modest sample sizes with diagnosed PSD meant statistical power was likely to be low. Therefore, secondly, PSD cases were combined with children presenting with a range of speech errors below the PSD threshold (N=141) and these children (N=404) were compared to the same control group (N=6399). These two derived PSD-related speech variables were used in analyses.

ALSPAC children were invited to clinics at the ages of 9.5 and 11.5 years, 7,725 and 7,159 children attended respectively. At these clinics, air conduction audiometry and tympanometry were used to assess hearing frequency thresholds and the function of the middle ear. Air conduction audiometry (at 11 years) and tympanometry (at 9 years) were used to assess the hearing and middle ear function

of ALSPAC children. The ear with the lowest threshold for each frequency was used to generate a measure of pure tone low frequency hearing by averaging across the 0.5, 1 and 2 kHz thresholds and a measure of pure tone high frequency hearing by averaging across the 3, 4 and 8 kHz thresholds [307,308]. In the tympanometry, middle ear pressure (MEP) was measured in both ears; the ear with the lowest MEP was used in analyses. Low MEP could suggest reduced Eustachian tube dysfunction, a risk factor for middle ear disease [308]. Variables pertaining to pure-tone high-frequency hearing, pure-tone low frequency hearing and MEP were used in analyses.

6.3.2 Polygenic risk score construction and analysis

6.3.2.1 PRS construction

As in **Chapter 5**, the most predictive PRS from **Chapter 3**, which includes LD-clumped SNPs with P-value< 0.0001 was used as a genetic proxy for nsCL/P. Using ALSPAC as a reference panel for LD, PLINK [161] was used to LD prune and clump the nsCL/P meta-analysis summary statistics (r^2<0.1 and 250 kb) using the most predictive P-value threshold. The PRS were then constructed in the ALSPAC sample. SNPs contained in the PRS are listed in **Table 16** in **Chapter 5.**

6.3.2.2 PRS power calculations

Power calculations for PRS analysis were performed using AVENGEME [136,224]. Initially, AVENGEME was used to estimate the minimum genetic covariance detectable with 80% power at P<0.05. Parameters included in calculations were the sample sizes of both the training and target samples, sample and population prevalences of nsCL/P, the number of SNPs common across both samples, a heritability estimate of nsCL/P and an estimate of the proportion of null markers that don't contribute to the disease risk of nsCL/P (**Table 25**).

Table 25: Parameters in PRS analysis power calculations

Parameter	Value	Source
Sample size of training sample (nsCL/P)	3987	determined from data
Sample size of target sample (Outcome)	Dependent on outcome	determined from data
Prevalence of nsCL/P in training sample	0.305	determined from data
Prevalence of outcome (if binary) in target sample	Dependent on outcome	determined from data
Population prevalence of nsCL/P	0.001	IPDTOC Working Group. "Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC)." (2011).
Population prevalence of outcome (if binary)	Dependent on outcome	determined from data
Number of independent SNPs common to both arrays (r^2<0.1)	75,737	determined from data
h^2 of nsCL/P	0.2	AVENGEME estimate in Chapter 3
Proportion of null markers	0.992	AVENGEME estimate in Chapter 3

However, it is difficult to directly compare the genetic covariance estimates between different phenotypes as the outcomes may differ in heritability. Therefore, as in **Chapter 5**, the genetic covariance estimates were transformed to genetic correlation estimates using the following formula:

$$Genetic\ Correlation_{AB} = \frac{Genetic\ Covariance_{AB}}{\sqrt{h_A^2 h_B^2}}$$

The AVENGEME heritability estimate of nsCL/P from **Chapter 3** was used for nsCL/P. GCTA [220] was used to estimate the heritability of adverse outcomes of interest. Where sample sizes were not sufficient to use GCTA (i.e. confidence

intervals overlap both 0 and 1), a SNP heritability of 0.2 for the related outcome was assumed for power calculations.

6.3.2.3 PRS association testing

The sample-size of ALSPAC children with both genotype and phenotype data varied across the different outcomes of interest from 661 to 5260. Therefore, power calculations were used to inform whether there was sufficient power to test the nsCL/P PRS for association with an outcome. Phenotypes with insufficient power (less than 80% power to detect a genetic correlation of 0.5) were removed from analyses.

For outcomes passing the power calculations, the associations between the nsCL/P PRS and the phenotypes were then estimated using a linear regression adjusted for the first four principal components. Effect sizes were reported per S.D. increase in PRS.

6.4 Results

6.4.1 Power calculations

Power calculations demonstrated that statistical power varied greatly across the different outcomes of interest. For anthropometric phenotypes, e.g. birthweight, there was sufficient power to detect small to moderate genetic correlation. Contrastingly, for dental outcomes there was insufficient power to detect even substantial genetic correlations. Worth nothing is that power calculations assumed a statistical significance threshold of $P<0.05$ and therefore do not consider statistical adjustment for the testing of multiple phenotypes which would further reduce power (**Table 26**).

Table 26: Power Calculations

Developmental outcome in ALSPAC	**Sample size with phenotype and genotype data**	**Minimum Genetic Covariance**	**SNP heritability estimate of outcome from GCTA (95% C.I.)**	**Minimum Genetic Correlation detectable (Using 95% C.I. from h^2 estimates)**
Birthweight	4887	0.039	0.24 (0.12, 0.36)	0.18 (0.15, 0.25)
Head circumference at birth	4887	0.039	0.23 (0.10, 0.35)	0.18 (0.15, 0.28)
Missing lateral incisor	724 controls 134 cases	0.150	N/A[2] (0.2)	0.75
Crossbite	585 controls 76 cases	0.185	N/A[2] (0.2)	0.93
PSD_1 / Controls	4606 controls 187 cases	0.095	N/A[2] (0.2)	0.48
Sub-clinical PSD_1 or PSD_1 / Controls	4606 controls 296 cases	0.082	N/A[2] (0.2)	0.41
Middle Ear Pressure	5260	0.038	0.07 [0, 0.19)	0.32 (0.19, 1]
High frequency hearing	5122	0.039	0.04 [0, 0.15)	0.44 (0.23, 1]
Low frequency hearing	5115	0.039	0.12 (0.00, 0.24)	0.25 (0.18, 1]

1 PSD: Persistent speech disorder

2 GCTA estimates were not reported if the confidence intervals overlapped both 0 and 1. A h2 of 0.2 was assumed for conversion to genetic correlation estimates.

6.4.2 The prediction of developmental phenotypes using PRS for nsCL/P

6.4.2.1 Anthropometric outcomes

There was little evidence for an association between the nsCL/P PRS and birthweight (Beta: -1.9g; 95% C.I. -12.6, 16.5; P = 0.79) or with head circumference at birth (Beta = -0.01 cm^2; 95% C.I. -0.05, 0.04; P =0.78). The confidence intervals suggested that there was unlikely to be substantial genetic correlation between these outcomes and nsCL/P (**Table 27**).

6.4.2.2 Speech and hearing outcomes

There was little evidence for an association between the nsCL/P PRS and persistent speech disorder (OR 1.05; 95% C.I. 0.90, 1.21; P=0.54). The result was consistent when sub-clinical cases were also included (OR 1.02; 95% C.I. 0.90,1.21 P=0.75). The wide confidence intervals suggest that the available data-set may be insufficient to rule out substantial genetic overlap (**Table 27**).

Similarly, no strong evidence was found of an association between the nsCL/P PRS and middle ear pressure at age 9 (Beta: 0.32 daPa; 95% C.I. -1.6, 2.3; P = 0.75), low frequency pure-tone hearing at age 11 (Beta: -1.9g; 95% C.I. -12.6, 16.5; P = 0.79) and high frequency pure-tone hearing at age 11 (Beta: -1.9g; 95% C.I. -12.6, 16.5; P = 0.79). Again, confidence intervals suggested that substantial genetic overlap was unlikely (**Table 27**).

Table 27: Association between Cleft PRS & developmental outcomes

Phenotype	**Increase in units per 1 SD increase in nsCL/P PRS (95% C.I.)**	**P value**
Anthropometric:		
Head circumference at birth (cm^2) N=4887	-0.01 (-0.05, 0.04)	0.78
Birthweight (g) N=4887	1.9 (-12.6, 16.5)	0.79
Speech:		
PSD diagnosis (OR compared to control) N=4793	1.05 (0.90, 1.21)	0.54
Sub-clinical PSD/PSD (OR compared to control) N=4902	1.02 (0.91, 1.15)	0.75
Hearing:		
Middle ear pressure at 9 years ($daPa_2$) N=5260	0.32 (-1.6, 2.3)	0.75
Low frequency hearing at 11 years (dB_3) N=5122	-0.06 (-0.20, 0.09)	0.43
High frequency hearing at 11 years (dB_3) N=5115	-0.06 (-0.27, 0.15)	0.58

1 PSD: Persistent speech disorder

2 daPa: decapascals a measure of air pressure

3 dB: decibels

6.5 Discussion

In this chapter, a nsCL/P PRS was found to be weakly associated with anthropometric measures, persistent speech disorder and hearing variables in a sample representative of the general UK population. The statistical power to detect genetic overlap varied across the different phenotypes; the large sample sizes and low effect sizes for the anthropometric and hearing variables (e.g. 1.9 grams or -0.06 decibels) suggest that for these outcomes, the true effect is unlikely to be of clinical significance. However, the wide confidence intervals for the speech variables suggested that strong conclusions cannot be made about genetic overlap between nsCL/P and speech.

In general, these findings imply that there are likely alternative causes of these adverse developmental outcomes in OFC cases distinct from genetic liability. Previously discussed possibilities, such as the physical presence of an OFC or complications of surgery, may be responsible for the differences in speech, hearing and anthropometric outcomes. For example, a CP often results in differences in the size and shape of the oral nasal and pharyngeal cavities and the velopharyngeal valve. These differences affect the physical production of sounds, which may increase the risk of speech disorders [309]. Similarly, the function of the Eustachian tube may be affected by the presence of a CP, which could lead to impaired hearing [110]. Lower birth weight in nsCL/P cases may be related to the increased risk of preterm birth in children with congenital malformations [310].

In **Chapter 5**, strong evidence was found for an aetiological link between nsCL/P and facial morphology, and previously identified sub-phenotypes in nsCL/P cases and unaffected relatives have been largely craniofacial phenotypes (lip pits, facial morphology, dental, orbicularis oris muscle defects) [93,96-98,102,298]. This

suggests that, of the outcomes considered in this chapter, dental outcomes may have been the most likely to be related to genetic liability to nsCL/P. Indeed, previous studies have found evidence of some genetic overlap [301]; *MSX1* is a risk locus for both nsCL/P and dental anomalies [65,302,311,312]. This is also supported by the genetic overlap between nsCL/P and OFC syndromes which can present with dental anomalies. For example, distinct genetic variation in *IRF6* is associated with nsCL/P and causes Van der Woude syndrome [93]. However, the available data were insufficiently powered to test the nsCL/P PRS for association with dental phenotypes.

There are several strengths of analyses described in this chapter. Firstly, the richness of ALSPAC as a data source; high quality phenotype data, for speech, anthropometric and hearing variables (which were derived by specialists), and genotype data were available for a large proportion of the cohort. Secondly, given that nsCL/P PRS were previously shown to be associated with philtrum width in **Chapter 5**, the method has been demonstrated to be effective at detecting genetic overlap and nsCL/P. Thirdly, the use of genetic methods to disentangle the possible causes of adverse outcomes in nsCL/P cases is a relatively unique design that bypasses the effects of surgery and the physical presence of an OFC.

Nevertheless, there are multiple limitations that need to be considered. Firstly, the heterogeneity of OFC subtypes complicates interpretation. In this study a PRS for nsCL/P was used but although nsCL/P is often treated as a single phenotype in epidemiological and genetic studies, distinct from CPO, there is emerging evidence that CLO and CLP have aetiological differences [8,254]. The phenotypic heterogeneity is a limitation as a CP may be the most relevant OFC to speech and hearing. In theory, a PRS from a combined grouping of CPO and CLP cases (nsCP/CL) may be

more predictive than the nsCL/P PRS. A second limitation is that statistical power was potentially low for speech-related phenotypes, meaning that modest to moderate effects cannot be ruled out. Thirdly, it is possible that low frequency or rare genetic variation not tagged by the nsCL/P PRS may be causally related to both nsCL/P and the increased risk of outcomes tested. However, this is unlikely to explain the consistent risk increase of adverse outcomes across all nsCL/P cases, and common genetic variation has been previously shown to play a major aetiological role for nsCL/P in both **Chapter 3** and in previous publications [24-28].

To conclude, weak evidence was found for genetic overlap between nsCL/P and speech, hearing and anthropometric outcomes. These results suggest that genetic liability to nsCL/P is unlikely to play a causal role in the increased risk of these adverse outcomes in children with nsCL/P. Therefore, these outcomes are likely to be related to if an individual has an OFC or related surgical procedures. Further research using larger samples are necessary to test the possibility of genetic overlap between nsCL/P and dental-related phenotypes.

Chapter 7: Estimating the genetic overlap between nsCL/P and oral cavity/ oropharyngeal cancer

7.1 Abstract

Observational studies suggest that individuals with birth defects, including nsCL/P, may have increased risk of cancer. An association between nsCL/P and cancer is further supported by several genes (e.g. *CDH1* and *AXIN2*) implicated in the aetiologies of both traits. However, the genetic overlap between nsCL/P and OC/OPC, which affect similar anatomical regions, has not been previously investigated.

nsCL/P PRS were constructed at different P-value inclusion thresholds in a sample of 5,048 OC/OPC cases and 5,450 controls of European ancestry. PRS analyses were followed up by MR analysis to test a possible causal relationship between liability to nsCL/P and risk of OC/OPC.

Strong evidence was found for an association between a 1 S.D. increase in nsCL/P PRS (including SNPs with P<0.1) and increased odds of OC/OPC (OR 1.09: 95% C.I. 1.04, 1.13 P = 5.3×10^{-5}). MR analyses found weak evidence for a causal effect of liability to nsCL/P on OC/OPC.

In follow-up analyses in the UK Biobank, weak evidence was found for an association between the nsCL/P PRS and tobacco smoking or alcohol behaviour phenotypes, two known risk factors for OC/OPC. Additionally, the association between the nsCL/P PRS and OC/OPC did not replicate in the UK Biobank sample of 687 cases and 408,282 controls (OR 1.01: 95% C.I. 0.93, 1.09 P=0.86). However, the lack of replication may be due to low statistical power, the confidence intervals

overlapped with the initial analyses. The implications of the results from the two studies are that nsCL/P and OC/OPC may have shared genetic aetiology, unrelated to alcohol or tobacco intake.

7.2 Introduction

Evidence from epidemiological population-based studies suggests that individuals with birth defects may have increased risk of developing cancer, particularly childhood cancers [313-316]. For example, individuals with chromosomal aneuploidies such as Down syndrome have been shown to have increased incidence of leukaemia [313-315]. However, the epidemiological evidence for increased incidence of cancer in nsCL/P cases is largely inconsistent. Some studies have reported increased incidence of cancer amongst OFC cases and unaffected relatives [123,313,317], but other studies have reported weak evidence for an association [122,314,315]. There are several limitations of an epidemiological approach for investigating the relationship between nsCL/P and cancer. Firstly, statistical power is limited by the modest number of individuals with both nsCL/P and cancer. Indeed one study reported findings from a data-set including 89 OFC cases, of which 2 had cancer [317]. Secondly, cancer is a highly heterogeneous phenotype; the aetiologies of childhood leukaemia and colorectal cancer are likely highly distinct, and subtype stratification would further reduce power [318,319]. Thirdly, depending on the available data, it can be difficult to differentiate between syndromic and non-syndromic OFCs, which have highly different aetiologies. For example, Patau syndrome [320], which can present with an OFC, is caused by trisomy, a known risk factor for leukaemia [313-315].

An alternative approach for disentangling the relationship between nsCL/P and cancer is to investigate their shared genetics; genetic overlap may suggest that nsCL/P cases have increased risk of cancer and potentially reveal common

biological pathways. Previous studies suggest that nsCL/P and cancer have shared genetic risk factors; several genes have been identified that are relevant to the aetiologies of both nsCL/P and cancer [127], notably *CDH1* [128,129,321] which is implicated in gastric and breast cancer [322], and *AXIN2* [129,323-325] which is linked to colorectal cancer and tooth agenesis [326] [327].

Cancers affecting the oral cavity and pharynx are a strong candidate for shared genetic aetiology with nsCL/P because both phenotypes affect similar anatomical sites. The main risk factors for OC/OPC relate to alcohol consumption, tobacco use or HPV infection [328]. However, OC/OPC has been shown to have a substantial genetic component, with 8 independent genetic risk loci identified in a previous GWAS [328] (although important to note that this includes a SNP in *ADH1B* known to be strongly associated with alcohol consumption [329]). Potential shared genetic aetiology between nsCL/P and OC/OPC has not been previously investigated, possible due to the relative rarity of both phenotypes.

As demonstrated in previous chapters, nsCL/P PRS can be used effectively to detect genetic overlap between nsCL/P and heritable phenotypes. Genetic overlap between nsCL/P and OC/OPC could suggest common risk factors or biological pathways involved in the aetiologies of both traits. A distinct possibility is that risk of OC/OPC is causally related to liability to nsCL/P. As in **Chapter 5**, the principles of MR can be used to test the possibility of this causal relationship.

To explore the shared genetics of nsCL/P and OC/OPC; firstly, nsCL/P PRS were used to explore potential genetic overlap between nsCL/P and OC/OPC. Secondly, in the instance of genetic overlap, MR was used to test a possible causal relationship between liability to nsCL/P and risk of OC/OPC. Thirdly, again in the

instance of genetic overlap, the association between nsCL/P PRS and OC/OPC was tested for replication in the UK Biobank and the PRS was investigated for association with known risk factors (alcohol and smoking intake) for OC/OPC. In this chapter I performed all described analyses.

7.3 Methods

7.3.1 Study participants

7.3.1.1 ICC and Bonn-II

The nsCL/P meta-analysis GWAS summary statistics, previously described in **Chapter 3**, were used to construct nsCL/P PRS. In brief, the meta-analysis summary statistics included ICC TDT results from 638 parent-offspring trios and 178 offspring duos of European descent, meta-analysed with GWAS summary results on 399 cases and 1,318 controls from the Bonn-II study.

7.3.1.2 OC/OPC GWAS data-set

The OC/OPC GWAS data-set [328] has been previous described in **Chapter 2**, including relevant information on how OC/OPC cases were identified (using ICD codes) and genotyping. In brief, the data-set consists of 12 epidemiological studies from North America, South America and Europe, comprising 6,034 OC/OPC cases and 6,585 controls. For the purposes of analyses, the data-set was restricted to 5,048 cases and 5,450 controls of recent European ancestry, which were split into studies from North America and studies from Europe.

7.3.1.3 UK Biobank

The UK Biobank has been previously described in **Chapter 2**, including information on genotyping. In brief, the UK Biobank is a large-scale cohort study of around 500,000 individuals, recruited from across the UK. For analyses in this

chapter, relevant phenotypes were OC/OPC case-control status and self-reported alcohol consumption/tobacco smoking variables.

ICD10 hospitalisation codes were used to identify OC/OPC cases in the UK Biobank using the same classification codes as in the OC/OPC GWAS data-set [328]. Oral cavity cancer (C02.0–C02.9, C03.0–C03.9, C04.0–C04.9, C05.0–C06.9), oropharyngeal (C01, C02.4, C09.0–C10.9), hypopharyngeal (C13.0–C13.9) and overlapping (C14). The number of cases was not sufficient for stratification by subtype, so all cases were grouped together.

Self-reported alcohol consumption data were collected at baseline using a questionnaire. Participants were asked for their alcohol drinking status (current, former, never) and for estimates of their average weekly intake of a range of different alcoholic beverages (red wine, white wine, champagne, beer, cider, spirits, fortified wine). From these variables, an average intake of alcoholic units per week was derived by summing the estimated intakes of the different alcoholic beverages consumptions across the seven drink types, as in a previous study [330]. The questionnaire used the following measurement units for each of the five alcoholic drink types: measures for spirits, glasses for wines and pints for beer/cider which were estimated to be equivalent to 1, 2 and 2.5 units respectively. Individuals reporting current intake frequency of “one to three times a month”, “special occasions only” or “never” (for whom this phenotype was not collected), were assumed to have a weekly alcohol consumption volume of 0.

Self-reported tobacco smoking data were also collected at baseline using a questionnaire. Participants were asked their tobacco smoking status; current, former, never. Current smokers were asked to estimate the number of cigarettes smoked per

day and the age they started smoking which was used to generate a pack years measure. Similarly, former smokers were asked to estimate the number of cigarettes smoked per day previously, the age they started smoking and the age they stopped smoking, to generate a pack years measure. Individuals reporting their tobacco smoking status as “never smokers” were assumed to have tobacco pack years of 0.

The mean and S.D. for weekly alcoholic consumption volume and tobacco pack years were calculated; individuals more than five S.D. away from the mean were removed from relevant analyses. Note that although under certain conditions (described above), an individual’s missing alcoholic volume and pack year phenotype data were assumed to be 0, the mean and the S.D. were calculated in the original sample reporting non-missing data.

7.3.2 *Polygenic risk score construction and analysis in OC/OPC data-set*

The 1000 Genomes (Phase 3) [163] CEU sample was used as a reference panel to LD clump (r^2<0.1 and 250 kb) the nsCL/P meta-analysis summary statistics at 11 different P-value inclusion thresholds (0.000001, 0.000005, 0.00001, 0.00004, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1). In previous chapters a single PRS threshold (P<0.00001) was used to test for genetic overlap with multiple phenotypes but given the focus on a single phenotype (and subtypes), it was decided to test a range of thresholds for a more comprehensive exploration of shared genetic aetiology.

Next, the nsCL/P PRS, at the 11 different P-value inclusion thresholds, were constructed separately in the European and North American OC/OPC case-control samples. The associations between the nsCL/P PRS and OC/OPC case-control status were then estimated in the two sub-samples for; all cases, oral cavity (OC)

cases only and oropharyngeal cases only (OPC) using logistic regression. For the comparison of OC/OPC subtypes, cases and controls with less than 70% CEU ancestry were removed. Covariates in the model were the first 10 genetic principal components, sex and age in both samples. METAL [241] was then used to meta-analyse the effect sizes, standard errors and P-values from the two sub-samples using a fixed-effects model.

7.3.3 Mendelian randomization analysis in OC/OPC data-set

In the instance of genetic overlap between nsCL/P and OC/OPC, Mendelian randomization (MR) was used to investigate a possible causal relationship between liability to nsCL/P and risk of OC/OPC. As in **Chapter 5**, six genome-wide significant SNPs for nsCL/P, taken from a previous publication [24], and listed in **Table 18** were used as a genetic instrument for liability to nsCL/P. The SNP-liability to nsCL/P information was extracted from the nsCL/P meta-analysis GWAS summary statistics. The European and North American OC/OPC GWAS summary statistics [328] were meta-analysed using METAL [241]. The relevant SNP-risk of OC/OPC information was extracted from this meta-analysis.

In a bidirectional MR approach, the reverse direction would also be tested (i.e. the effect of liability to OC/OPC on risk of nsCL/P). However, of the eight genome-wide significant SNPs, four were specific to oral cavity cancer, a further two had obvious issues with pleiotropy (a SNP in *ADH1B* associated with alcohol consumption and a SNP in the *HLA* region) and one variant was a rare intronic deletion. The reverse direction MR analysis was not performed because of the limitations of available genetic instruments for OC/OPC.

7.3.4 Testing PRS for replication in the UK Biobank and for association with environmental risk factors

The UK Biobank was used to follow-up results from the OC/OPC GWAS data-set to firstly, attempt to replicate results and secondly, to investigate the association of nsCL/P PRS with known environmental risk factors for OC/OPC. Firstly, the UK Biobank data-set of 487,409 individuals was restricted to a subset of 409,700 individuals of recent European descent. Individuals of non-European descent were removed based on a k-means cluster analysis on the first 4 genetic principal components [331].

Secondly, the nsCL/P PRS most strongly associated with case-control status from analyses described in **Chapter 7.3.2** was constructed in the sample. Finally, the nsCL/P PRS was tested for association with OC/OPC case-control status, alcoholic units consumed per week and tobacco smoking pack years adjusting for sex, age and the first 10 genetic principal components.

7.4 Results

7.4.1 The prediction of OC/OPC risk using PRS for nsCL/P

After meta-analysing the North American and European results, strong evidence was found for an association between nsCL/P PRS and increased risk of OC/OPC. A 1 S.D. increase in nsCL/P PRS (including independent SNPs with P<0.1) was associated with increased odds of OC/OPC (OR 1.09; 95% C.I. 1.04, 1.13; P=0.000053). PRS with more liberal inclusion thresholds (e.g. P < 0.05 and P < 0.1) which included thousands of SNPs were more strongly associated with risk of OC/OPC than more conservative inclusion thresholds (**Table 28**).

Table 28: Association of nsCL/P PRS with risk of OC/OPC

Polygenic risk score P-value inclusion threshold	Number of SNPs in PRS	All OC/OPC subtypes against controls (5048 Cases and 5450 Controls)	
		OR (95% C.I.) Per 1 S.D. increase in PRS	P
0.000001	10	1.01 (0.97, 1.05)	0.64
0.000005	15	1.02 (0.98, 1.07)	0.27
0.00001	18	1.02 (0.98, 1.07)	0.27
0.00005	48	1.01 (0.97, 1.05)	0.71
0.0001	78	1.01 (0.97, 1.05)	0.58
0.0005	238	1.01 (0.97, 1.05)	0.71
0.001	424	1.01 (0.96, 1.05)	0.79
0.005	1,607	1.05 (1.01, 1.09)	**0.021**
0.01	2,777	1.03 (0.99, 1.08)	0.10
0.05	8,620	1.06 (1.02, 1.11)	**0.0026**
0.1	12,614	1.09 (1.04, 1.13)	**0.000053**

Similarly, there was strong evidence for an association between the nsCL/P PRS and the two main OC/OPC subtypes, oropharyngeal and oral cavity cancer. A 1 S.D. increase in nsCL/P PRS (P<0.1) was associated with increased odds of both oropharyngeal cancer (OR 1.10; 95% C.I. 1.04, 1.16; P=0.00079) and oral cavity cancer (OR 1.12; 95% C.I. 1.06, 1.17; P=0.000016) (**Table 29**).

Table 29: Association of nsCL/P PRS with OC/OPC subtypes

Polygenic risk score P-value inclusion threshold	Oropharyngeal cases only (2297 Cases and 5182 Controls)		Oral cavity cases only (2463 Cases and 5182 Controls)	
	OR (95% C.I.) Per 1 S.D. increase in PRS	P	OR (95% C.I.) Per 1 S.D. increase in PRS	P
0.000001	1.03 (0.97, 1.08)	0.37	1.01 (0.96, 1.07)	0.58
0.000005	1.04 (0.98, 1.09)	0.19	1.02 (0.97, 1.07)	0.54
0.00001	1.03 (0.98, 1.09)	0.25	1.02 (0.97, 1.07)	0.44
0.00005	1.02 (0.96, 1.07)	0.51	1.02 (0.97, 1.08)	0.35
0.0001	1.03 (0.97, 1.08)	0.38	1.03 (0.98, 1.09)	0.20
0.0005	1.01 (0.96, 1.07)	0.62	1.03 (0.98, 1.08)	0.23
0.001	1.02 (0.97, 1.08)	0.45	1.02 (0.97, 1.08)	0.35
0.005	1.04 (0.99, 1.10)	0.11	1.07 (1.02, 1.13)	**0.0061**
0.01	1.04 (0.99, 1.10)	0.12	1.05 (1.00, 1.10)	0.073
0.05	1.07 (1.02, 1.13)	**0.011**	1.09 (1.04, 1.15)	**0.00053**
0.1	1.10 (1.04, 1.16)	**0.00079**	1.12 (1.06, 1.17)	**0.000016**

7.4.2 Mendelian randomization

The follow-up MR analysis testing the causal effect of liability to nsCL/P on risk of OC/OPC (using the 6 SNPs previously used in **Chapter 5** and listed in **Table 18**) found weak evidence for a causal relationship. A 1-unit log odd increase in liability to nsCL/P was weakly associated with decreased odds of OC/OPC (OR 0.98; 95% C.I. 0.91, 1.05; P = 0.53). MR sensitivity analyses generated similar effect estimates (**Table 30**).

Table 30: MR analysis of liability to nsCL/P on OC/OPC risk

Test	Interpretation	OR (95% C.I.)	P-value
Inverse variance weighted	Primary result[1]	0.98 (0.91, 1.05)	0.53
Heterogeneity of Inverse variance weighted	Balanced pleiotropy	N/A	0.37
MR-Egger	Intercept test for directional pleiotropy[2]	0.03 (-0.06, 0.11)	0.60
	Regression estimate[1]	0.93 (0.77, 1.13)	0.49
Weighted median	Consistency[1]	0.96 (0.90, 1.03)	0.30
Weighted mode	Consistency[1]	0.95 (0.88, 1.03)	0.30

1 Units: Odds ratio for OC/OPC per 1-unit log odd increase in liability to nsCL/P

2 Units: Average pleiotropic effect of a nsCL/P genetic variant on odds of OC/OPC

7.4.3 Testing PRS for replication in the UK Biobank and for association with environmental risk factors

In the follow-up analysis in the UK Biobank, weak evidence was found that the nsCL/P PRS (P<0.1), previously found to be associated with OC/OPC in the GWAS data-set, is associated with OC/OPC (OR 1.01; 95% C.I. 0.93,1.09; P=0.85). However, this may be because of the smaller sample sizes in the UK Biobank (687 cases compared to 5,048). Indeed, the confidence intervals overlapped with the previous estimate. Additionally, there was no strong evidence for an association between the nsCL/P PRS and self-reported alcohol consumption or lifetime cigarette smoking. The confidence intervals for alcoholic units per week (-0.09 to 0.03 alcoholic units per week) and pack years (-0.08 to 0.02 pack years) suggested any true effect, if one exists, is modest and unlikely to explain the association between the nsCL/P PRS and OC/OPC observed in the OC/OPC GWAS sample (**Table 31**).

Table 31: Association of nsCL/P PRS (P<0.1) with OC/OPC, alcohol and smoking in the UK Biobank

Phenotype (Units)	**N**	**Effect size (95% C.I.)** *Per 1 S.D. increase in nsCL/P PRS*	**P-value**
OC/OPC case-control status (OR)	687 cases 408,282 controls	1.01 (0.93, 1.09)	0.85
Alcohol consumption (Units per week)	291,944	-0.03 (-0.09, 0.03)	0.34
Lifetime cigarette smoking (Pack years)	123,685	-0.03 (-0.08, 0.02)	0.31

7.5 Discussion

In this chapter, nsCL/P PRS were found to be associated with increased risk of OC/OPC, with concordant results from subtype analyses, but follow-up MR analyses found weak evidence for a causal effect of liability to nsCL/P on risk of OC/OPC. These results imply that the most likely possibility is that nsCL/P and OC/OPC have shared genetic aetiology. Although the relationship is non-causal, genetic overlap suggests that individuals with nsCL/P and other individuals with high genetic liability to nsCL/P (e.g. unaffected family members) may have slightly elevated risk of developing OC/OPC because of shared genetic risk factors.

However, the specific biological pathways that the nsCL/P PRS is proxying for, relevant to both nsCL/P and OC/OPC, are currently unclear. Well-powered analyses in the UK Biobank suggested that alcohol and cigarette smoking behaviour are unlikely to explain the genetic overlap. The interpretation of what exactly a nsCL/P PRS is tagging becomes increasingly complex as more and more SNPs are included. Indeed, the nsCL/P PRS most strongly associated with OC/OPC included over 10,000 SNPs. One possibility is that the genetic overlap may be attributable to genes involved both in early development and tumour suppression. For example, the

CDH1 gene, known to be related to nsCL/P and cancer subtypes, has been shown to be related to both axonal growth and patterning in the developing murine brain [332], and tumour suppression [333].

Another possibility is that, although the maternal environment is the route to exposure for nsCL/P, the nsCL/P PRS may be proxying for environmental exposures. The maternal genotype is correlated with the foetal genotype suggesting that the nsCL/P PRS may be a weak proxy for maternal genetic risk factors. Furthermore, both the foetal and maternal genotype may both play important roles in the effect of environmental exposures on risk of OFCs. For example, there is some evidence that the maternal and foetal *ADH1C* haplotype may modify the association between maternal alcohol consumption and risk of OFCs via alcohol metabolism [334].

Previous epidemiological and genetic studies have not investigated the relationship between nsCL/P and OC/OPC. However, the findings in this chapter are consistent with previous genetic studies that have identified genetic variation associated with nsCL/P and various cancer subtypes [127-129,321,323-325]. It is difficult to compare the results to the epidemiological literature as the majority of studies focused on childhood cancers but the results are consistent with the reported effect size from a previous study, investigating the risk of adult-onset cancers in nsCL/P cases, which was likely underpowered to detect a modest effect [122].

There are several strengths of analyses in this chapter. The usage of PRS as a genetic proxy for nsCL/P has advantages over the candidate gene or candidate SNP approaches used in previous studies [127-129,321,323-325] because it both reduced the number of statistical tests, and more generally, extended the evidence for genetic overlap at specific loci to evidence of genome-wide genetic overlap. A further

strength is the focus on OC/OPC cancer; previous epidemiological and genetic studies have looked more generally at the relationship between nsCL/P and all cancers but cancers arising from different organs may be highly aetiologically heterogeneous. Although similar results were found in analyses for the OC and OP subtype analyses, it is important to note that these subtypes may also be aetiologically heterogeneous [328]. Finally, the study design is a considerable strength; the construction of nsCL/P PRS derived from a modestly sized GWAS in a much larger OC/OPC GWAS data meant that analyses were well-powered to detect modest genetic overlap. Similarly, the use of the UK Biobank cohort allowed well-powered exploration of possible biological mechanisms.

Nevertheless, there are several limitations of analyses. Firstly, the absence of replication of the association between the nsCL/P PRS and OC/OPC in the UK Biobank weakens the argument for genetic overlap between the two phenotypes. However, this may be because of the modest number of OC/OPC cases in the UK Biobank, confidence intervals overlapped with the estimates from the OC/OPC GWAS data-set. Secondly, the OC/OPC GWAS data-set was highly ancestrally heterogeneous and included samples from 12 different epidemiological studies, including a case-only study (the Head and Neck 5000). Therefore, it is possible that allele frequency differences between cases and controls relating to population differences could result in spurious associations with the nsCL/P PRS, although this is unlikely given the number of SNPs in the PRS. Thirdly, available genetic instruments were insufficient to test the hypothesis that liability to OC/OPC is associated with risk of nsCL/P, so this possibility cannot be ruled out. Finally, as in previous chapters, treating heterogeneous nsCL/P subtypes (CL/P, CLO) [8,9,27] as a single phenotype, complicates the interpretation of results. Considering anatomical

site, the oropharynx includes the soft palate while the oral cavity includes the lips and hard palate, suggesting that the different nsCL/P subtypes may have different mechanistic relationships with the OC/OPC subtypes.

To conclude, nsCL/P and OC/OPC likely have shared genetic risk factors that are unrelated to alcohol or tobacco intake. nsCL/P cases and unaffected family members may also have slightly increased risk of OC/OPC. Further work is required to more formally investigate common biological pathways and shared environmental risk factors between nsCL/P and OC/OPC.

Chapter 8: Discussion

8.1 General discussion

A recurring theme throughout this thesis is the use of PRS and/or MR to explore the causes and consequences of nsCL/P. The modest sample sizes of the available nsCL/P GWAS data meant that, in general, analyses investigating liability to nsCL/P as an exposure, in conjunction with larger outcome data-sets, were more statistically viable than analyses considering nsCL/P as the outcome. For this reason, the hypothesis that the inheritance of many binary traits such as nsCL/P is determined by a continuous, normally distributed variable, i.e. liability, was integral to many analyses. An interesting possibility is that an individual's underlying liability to nsCL/P could have causal effects on nsCL/P-related phenotypes in individuals without nsCL/P.

The liability threshold model was proposed as an explanation for the mode of inheritance of binary traits that did not follow Mendelian inheritance patterns [156,202,293,335]. Indeed, Carter first proposed a multifactorial model of inheritance for nsCL/P in the late 1960's [202]. However, geneticists of the past did not have access to the abundance of genotype data, which are now currently available, and so were limited by their reliance on pedigree-recurrence rates methods. Indeed, several pedigree studies incorrectly concluded that the segregation patterns of nsCL/P were inconsistent with a multifactorial model of inheritance [207-210], which may have been the reason that linkage analysis was utilised to map a theoretical major nsCL/P gene [17-20,211]. The difficulties mapping a major nsCL/P susceptibility gene using linkage were followed by the subsequent identification of many independent nsCL/P risk loci in association studies [9,21-24,27-30,218], and a multifactorial model of inheritance for nsCL/P is now widely accepted [13].

In **Chapter 3**, I extended previous work on the genetics of nsCL/P by using several different genotype-driven methods to further explore the genetic architecture of nsCL/P and estimate the SNP heritability. Limitations of the available data-sets and case-control matching meant that GCTA SNP heritability estimates were inflated but estimates from other methods such as AVENGEME and LD-score regression were less susceptible to these limitations. Triangulating results from the different methods suggested, firstly; that nsCL/P has a SNP heritability of between 20 and 33%, highly consistent with a published GCTA estimate of 30% [26] and secondly; that nsCL/P has a highly polygenic architecture. These findings supported the use of nsCL/P PRS in subsequent chapters, to detect genetic overlap between nsCL/P and phenotypes of interest. Furthermore, the abundance of well-characterised nsCL/P risk loci with large effect sizes, suggested that an MR framework could be used effectively to test causal effects of liability to nsCL/P. The premise that liability to nsCL/P could have phenotypic effects in the general population was tested in the four subsequent results chapters.

In **Chapter 4**, MR and joint-likelihood co-localisation were used to test the hypothesis that the effect of nsCL/P genetic risk variants on disease liability is mediated by DNA methylation. Three putative loci at *VAX1*, *LOC146880* and *NTN1* were identified where genetic variation may affect nsCL/P via DNA methylation. Functional follow-up found that two of the loci were also associated with gene expression of nearby genes. However, there were several limitations meaning that identified loci are merely putative. Firstly, the methylation GWAS data-set was even smaller than the nsCL/P GWAS data-set, meaning MR analyses were insufficiently powered to test the causal effect of liability to nsCL/P on methylation. Secondly, with the available genetic instruments, it was not possible to distinguish between vertical

and horizontal pleiotropy. Thirdly, although results were compared with DNA methylation data derived from lip and palate, primary analyses used DNA methylation in blood as a proxy for more relevant tissues. Nevertheless, the work in this chapter extended previous work investigating the role of DNA methylation in nsCL/P [8,34-36]. The use of a previously-devised bioinformatics-based framework [267,268], publicly available GWAS summary data and the principles of MR, allowed a greater focus on causal relationships between DNA methylation and nsCL/P than previous studies.

The focus of **Chapter 5** was unravelling the shared genetics between nsCL/P and normal-range variation in facial morphology. Starting with seven facial phenotypes, the width of the philtrum was found to be the most relevant facial phenotype to nsCL/P. nsCL/P PRS were found to be strongly predictive of philtrum width in the general population, across around 6,000 individuals of European ancestry from two studies. Follow-up MR analyses suggested that increased genetic liability to nsCL/P results in decreased philtrum width. These findings suggested that the inheritance of nsCL/P may be related to the additive effect of nsCL/P genetic risk variants on the development of the philtrum. Furthermore, these results demonstrated that genetic liability to nsCL/P has phenotypic effects in individuals without nsCL/P, validating the use of similar methods in later chapters.

Chapter 6 focused on potential genetic overlap between nsCL/P and a range of phenotypes for which nsCL/P cases are thought to have increased risk of adverse outcomes than controls, such as speech and hearing. The idea was to determine if the adverse outcomes are caused by the binary status of having nsCL/P, i.e. related to surgery or the physical presence of a cleft, or are on a continuum, related to liability to nsCL/P. Across all outcomes tested, weak evidence was found for shared

genetic effects between nsCL/P and adverse outcome phenotypes, but some analyses may have been limited by statistical power. The lack of genetic overlap suggests that the adverse outcomes tested are more likely to be related to complications of having an OFC rather than to an individual's genetic liability to nsCL/P.

Chapter 7 used a similar approach to disentangle possible genetic overlap between nsCL/P and cancer. Cancers can be highly heterogeneous so the relationship between nsCL/P and a grouping of cancers affecting a similar anatomical region, OC/OPC, was investigated. nsCL/P PRS including thousands of SNPs were found to predict increased risk of OC/OPC but MR analyses suggested that the increased risk is unrelated to genetic liability to nsCL/P. Follow-up analyses found weak evidence that the genetic overlap was related to smoking or tobacco use, and the exact reason for the shared genetic overlap remains unclear. The implications of these findings are that nsCL/P and OC/OPC have shared genetic aetiology that could be related to shared environmental risk factors or common biological processes.

The five results chapters investigated the same phenotype and used very similar methods but attempted to answer very different questions. nsCL/P was shown to be highly polygenic with a substantial SNP heritability, DNA methylation and the philtrum were both shown to have aetiological relevance to nsCL/P, nsCL/P and a cancer subtype were shown to have genetic overlap while contrastingly, weak evidence was found for genetic overlap between nsCL/P and several developmental outcomes that are common in nsCL/P cases.

8.2 Future work

8.2.1 Extensions to my thesis work

There are many possible extensions to the work in this thesis. Firstly, relevant to the work in **Chapter 3**; GWAS have been very effective at picking up high frequency genetic risk variants for nsCL/P with modest sample sizes but the contribution of low frequency and rare variants to the aetiology of nsCL/P is less clear. Increased sample sizes and methodological advances, such as imputation and sequencing, will allow increased opportunity to identify lower frequency risk variation for nsCL/P.

A recurring limitation throughout analyses in this thesis is the phenotypic heterogeneity of nsCL/P. As discussed previously, there is a growing body of evidence suggesting aetiological differences between nsCL/P subtypes (CLO and CLP) [8,28,254]. As the sample sizes of nsCL/P GWAS increase, there will be more power to test for subtype differences for genetic risk factors. Unravelling the genetics of CPO, which is not included within the nsCL/P subtype, and has an incidence of around 1 in 2,500 [336] is also important. CPO has been previously shown to have almost no polygenic overlap with nsCL/P and is thought to be more related to rare genetic variation [26]. To date, a single functional missense variant has been identified in a relatively small GWAS [336], larger efforts may identify additional risk loci.

Although the functional relevance of some nsCL/P risk variants is known, this is not the case for many identified loci. In **Chapter 4**, I investigated the possibility of genetic risk variants acting via DNA methylation. As DNA methylation, gene expression and other omics databases increase in size there is increased potential to use bioinformatics approaches to unravel biological pathways and improve aetiological understanding. However, investigating the epigenetics of nsCL/P is

difficult because the relevant tissue cannot be extracted during development, when an OFC forms. The increased sampling of more aetiologically relevant tissues such as lip and palate tissues over blood will improve the design of studies investigating the role of epigenetic mechanisms in the aetiology of nsCL/P.

An argument could be made that investigating the genetic architecture of other congenital multifactorial defects such as pyloric stenosis, a gastrointestinal defect leading to severe vomiting, and neural tube defects, defects of the spinal cord, could lead to improved understanding of congenital malformations in general, including nsCL/P. This is because of the possibility that different congenital defects may share genetic risk factors and involve common developmental pathways. A previous modestly sized GWAS [337] identified three risk loci for pyloric stenosis but there have not been large genetic studies for many other congenital defects likely because of the rarity of many multifactorial birth defects. Future work could investigate the possibility of genetic overlap between different congenital defects, although recruiting the necessary samples is likely to be a challenge.

In **Chapter 5**, a relationship between liability to nsCL/P and philtrum width was demonstrated but genetic overlap between nsCL/P and other facial phenotypes is also possible, including phenotypes tested in my analyses. Larger sample sizes and improved facial phenotyping, such as the recent global and local facial variation measures may allow further investigation of the shared genetics of nsCL/P and facial morphology [338].

The relationship between genetic liability to nsCL/P and philtrum width highlighted the potential aetiological relevance of the philtrum to the development of a CL. An extension could be to investigate the relationship between nsCL/P PRS

and the severity of an OFC, a dose-response effect would further support the liability threshold model.

Of the adverse outcomes considered in **Chapter 6**, dental anomalies were possibly the most biologically plausible to have genetic overlap with nsCL/P but there wasn't sufficient power to test for an association. Combining genetic data on dental anomalies from multiple cohorts would allow for this potential relationship to be investigated in future studies.

In **Chapter 7**, genetic overlap between nsCL/P and OC/OPC was demonstrated but the shared genetic risk factors are currently unclear. Further work could investigate common biological processes or shared environmental risk factors between the two phenotypes. Additionally, further work is necessary to determine if there is genetic overlap between nsCL/P and cancers of other anatomical regions, or if the genetic overlap is specific to OC/OPC. It is possible that some genetic risk variants are shared by nsCL/P and all cancer subtypes, which would suggest a general link between nsCL/P and cancer. Alternatively, if the genetic overlap is specific to OC/OPC this would suggest more localised tissue specific effects. If strong evidence is found of genetic overlap between nsCL/P and all cancer subtypes this could be translated to additional follow-up of nsCL/P cases.

8.2.2 Maternal environmental risk factors for nsCL/P

A planned thesis analysis was to use a MR framework to investigate possible maternal environmental risk factors for nsCL/P, identified from the observational epidemiological literature. However, there were several notable limitations of using the available data-sets to infer causality of maternal risk factors using MR which meant that the analysis was not performed. Below, I discuss the limitations of my

data-set, the importance of using MR to investigate maternal risk factors for nsCL/P, the theoretical required data-sets and further considerations

8.2.2.1 Limitations of the ICC data-set for investigation of maternal environmental risk factors

A major limitation was that the ICC data-set, consisting of parent-offspring trios, was missing a suitable control group for comparison with the mothers of cases. Matching to population controls and using paternal controls were both considered but these methods were potentially susceptible to a range of biases including, population stratification, batch effects, a causal effect of offspring genotype and assortative mating [339]. In particular, the difficulties of matching across different studies and genotyping chips were demonstrated in **Chapter 3**.

A second limitation was that many of the exposures of interest from the epidemiological literature have only one or two genetic risk variants identified at genome-wide significance (e.g. zinc [340]). Relaxing the inclusion threshold and using GWAS summary statistics to construct allele scores would be one possibility to increase power [341]. Another possibility would be to use the MR-RAPS estimator, a profile score based method which benefits from the inclusion of many SNPs, even weakly associated SNPs [342]. However, more than half of the ICC sample was non-European, and there is an absence of GWAS data for many of the exposures of interest in non-European populations. The use of genetic instruments derived in Europeans in an Asian population, without appropriate testing, could reduce power or even bias causal estimates.

A final limitation is that even in the best case scenario, MR power calculations [343] demonstrated that the available sample sizes are insufficient to detect modest to moderate causal effects. Assuming that both, genetic instruments explaining 5% of

the variation in both European and Asian populations are available, and each nsCL/P mother could be successfully matched to 4 controls, the data-set has 80% power to detect an OR of 1.38 (**Table 32**). Previous meta-analyses of epidemiological studies suggested that for possible maternal risk factors for nsCL/P, the magnitude of effect (if there is a causal effect) is likely to be small to moderate [57,67].

Table 32: Power calculations for the four possible study designs

Proportion of variance explained in exposure by genetic instruments (R^2)	OR per 1 S.D. increase in exposure detectable with 80% power			
	745 European mothers and 2980 matched controls	745 European mothers and 659 paternal controls	1598 European and Asian mothers and 6392 matched controls	1598 European and Asian mothers and 1481 paternal controls
0.01	2.31	4.26	1.88	2.69
0.02	1.91	2.85	1.61	2.03
0.03	1.73	2.36	1.49	1.79
0.04	1.63	2.11	1.42	1.66
0.05	1.56	1.95	1.38	1.57

8.2.2.2 Importance of MR to test maternal risk factors for nsCL/P?

Given that the available data-sets have insufficient sample sizes for MR analyses of environmental risk factors for nsCL/P, it is first important to consider the importance of identifying maternal risk factors for nsCL/P and the advantages of specifically using MR.

Importance is highly subjective, but from an epidemiological perspective, direct translation of MR results for prevention of nsCL/P is likely to be difficult given the probable modest effects of exposures and the relative rarity of nsCL/P. However, there are two important reasons for determining the causality of maternal exposures

for nsCL/P. Firstly, determining causal relationships and exploring the biological mechanisms can lead to increased aetiological understanding of nsCL/P and potentially other congenital anomalies. Secondly, if an exposure that is already known to have intra-uterine effects (e.g. maternal folate) is shown to be aetiologically relevant to nsCL/P, this would strengthen the argument for public health interventions on folate levels during pregnancy.

The use of MR to evaluate the potential causality of maternal exposures for nsCL/P may have advantages over both observational epidemiological studies and candidate gene approaches (e.g. using a single variant in *MTHFR*, a gene implicated in folate metabolism). Furthermore, comparing and contrasting the evidence from different methodologies, which have different sources of bias, can help to strengthen conclusions [344].

8.2.2.3 Required data-sets?

The next consideration is the required data-set for well-powered and non-biased testing for MR analyses. A major challenge is recruitment because the ideal data-set would have genetic data for affected offspring, both of their parents and the mothers of controls [339]. Power calculations suggest that even with relatively strong genetic instruments, much larger datasets than the ICC are required for MR analyses to be well-powered to detect these effect sizes (**Table 33**).

Table 33: What sample sizes are required to investigate maternal risk factors for nsCL/P?

True OR per 1 S.D. increase in exposure	Proportion of variance explained in exposure by genetic instruments (R^2)	Number of cases required (Assuming 1 case paired with 4 controls)
1.1	0.01	101,918 nsCL/P mothers
	0.02	50,959 nsCL/P mothers
	0.03	33,973 nsCL/P mothers
	0.04	25,480 nsCL/P mothers
	0.05	20,384 nsCL/P mothers
1.2	0.01	26,373 nsCL/P mothers
	0.02	13,187 nsCL/P mothers
	0.03	8,791 nsCL/P mothers
	0.04	6,594 nsCL/P mothers
	0.05	5,275 nsCL/P mothers
1.3	0.01	12,092 nsCL/P mothers
	0.02	6,046 nsCL/P mothers
	0.03	4,031 nsCL/P mothers
	0.04	3,023 nsCL/P mothers
	0.05	2,419 nsCL/P mothers

Assembling samples of these sizes would be a substantial undertaking as OFCs are relatively uncommon; with an estimated incidence of 1 in 700 for all non-syndromic OFCs and 1 in a 1000 for nsCL/P [1]. The need for such large study samples could be alleviated by the derivation of stronger genetic instruments for the maternal exposures. The use of non-specific allele scores consisting of thousands of variants, not necessarily reaching genome-wide significance, could further increase power [341]. However, whether genetic instruments explaining more than 5% of the variation in an exposure can be derived for exposures of interest is largely dependent on the genetic architecture and heritability of the exposure, and therefore may not be possible in some instances.

8.2.2.4 Further considerations

Assuming the availability of appropriate data-sets, there are several further complications that should be considered. Firstly, a potential concern is that a

proposed maternal exposure may also increase the risk of miscarriage. If children with nsCL/P have increased frailty (and so are at increased risk of miscarriage) and an exposure is associated with miscarriage, then there is potential for survival bias. Secondly, depending on the study population, gene-environmental interactions may distort the interpretation. For example, in the US, folate fortification is now mandatory and so in an MR analysis, the true effect of folate on nsCL/P may be underestimated. Another example is that if alcohol behaviour during pregnancy differs across different cultural groups, the maternal alcohol SNPs in an MR analysis may not be associated with expected outcomes in the offspring (i.e. alcohol SNPs will not be associated with outcomes in groups who do not drink alcohol for cultural reasons). Thirdly, many of the proposed risk factors may have complex interlinked relationships. Both, maternal smoking and adiposity have been explored as risk factors but these relationships are complicated by the fact that smoking is known to cause weight loss [345]. Similarly, disentangling the relationship between the different phenotypes in the one-carbon metabolism folate pathway may require more complex modelling. Multivariable MR could be one approach used to tackle this issue [346].

8.3 Summary

The results presented in this thesis are a starting point for the use of PRS and MR to unravel the causes and consequences of nsCL/P. Notable successes of this preliminary work included demonstration of shared genetics between nsCL/P and the philtrum, evidence for DNA methylation mediating the effects of nsCL/P genetic risk variants and evidence of genetic overlap between nsCL/P and a cancer subtype. Future work, with larger sample sizes, will have improved power to explore the aetiology of nsCL/P, unravel subtype heterogeneity and test the phenotypic consequences of genetic liability to nsCL/P.

References

1. Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *The Lancet.* 2009;374(9703):1773-1785.
2. Som P, Naidich T. Illustrated review of the embryology and development of the facial region, part 1: early face and lateral nasal cavities. *American Journal of Neuroradiology.* 2013;34(12):2233-2240.
3. Jiang R, Bush JO, Lidral AC. Development of the upper lip: morphogenetic and molecular mechanisms. *Developmental Dynamics.* 2006;235(5):1152-1166.
4. Som P, Naidich T. Illustrated review of the embryology and development of the facial region, part 2: late development of the fetal face and changes in the face from the newborn to adulthood. *American Journal of Neuroradiology.* 2014;35(1):10-18.
5. Larsen WJ, Sherman LS. *Human Embryology.* Vol 1: Churchill Livingstone New York; 1993.
6. Tolarová MM, Cervenka J. Classification and birth prevalence of orofacial clefts. *American Journal of Medical Genetics Part A.* 1998;75(2):126-137.
7. Murray J. Gene/environment causes of cleft lip and/or palate. *Clinical Genetics.* 2002;61(4):248-256.
8. Sharp GC, Ho K, Davies A, et al. Distinct DNA methylation profiles in subtypes of orofacial cleft. *Clinical Epigenetics.* 2017;9(1):63.
9. Leslie EJ, Carlson JC, Shaffer JR, et al. Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Human Genetics.* 2017;136(3):275-286.
10. Kondo S, Schutte BC, Richardson RJ, et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature Genetics.* 2002;32(2):285-289.
11. Malik S, Kakar N, Hasnain S, Ahmad J, Wilcox E, Naz S. Epidemiology of Van der Woude syndrome from mutational analyses in affected patients from Pakistan. *Clinical Genetics.* 2010;78(3):247-256.
12. Venkatesh R. Syndromes and anomalies associated with cleft. *Indian Journal of Plastic Surgery.* 2009;42(3):51.
13. Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. *Nature Reviews Genetics.* 2011;12(3):167-178.
14. Terwilliger JD, Ott J. *Handbook of Human Genetic Linkage.* JHU Press; 1994.
15. Murray J, Nishimura D, Buetow K, et al. Linkage of an autosomal dominant clefting syndrome (Van der Woude) to loci on chromosome Iq. *American Journal of Human Genetics.* 1990;46(3):486.
16. Lees MM, Winter RM, Malcolm S, Saal HM, Chitty L. Popliteal pterygium syndrome: a clinical study of three families and report of linkage to the Van der Woude syndrome locus on 1q32. *Journal of Medical Genetics.* 1999;36(12):888-892.
17. Eiberg H, Bixler D, Nielsen L, Conneally P, Mohr J. Suggestion of linkage of a major locus for nonsyndromic orofacial cleft with F13A and tentative assignment to chromosome 6. *Clinical Genetics.* 1987;32(2):129-132.
18. Scapoli L, Pezzetti F, Carinci F, Martinelli M, Carinci P, Tognon M. Evidence of linkage to 6p23 and genetic heterogeneity in nonsyndromic cleft lip with or without cleft palate. *Genomics.* 1997;43(2):216-220.

19. Stein J, Mulliken JB, Stal S, et al. Nonsyndromic cleft lip with or without cleft palate: evidence of linkage to BCL3 in 17 multigenerational families. *American Journal of Human Genetics.* 1995;57(2):257.
20. Hecht JT, Wang Y, Connor B, Blanton SH, Daiger SP. Nonsyndromic cleft lip and palate: no evidence of linkage to HLA or factor 13A. *American Journal of Human Genetics.* 1993;52(6):1230.
21. Beaty TH, Murray JC, Marazita ML, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genetics.* 2010;42(6):525-529.
22. Birnbaum S, Ludwig KU, Reutter H, et al. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genetics.* 2009;41(4):473-477.
23. Mangold E, Ludwig KU, Birnbaum S, et al. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genetics.* 2010;42(1):24-26.
24. Ludwig KU, Mangold E, Herms S, et al. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nature Genetics.* 2012;44(9):968-971.
25. Leslie EJ, Carlson JC, Shaffer JR, et al. A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24. 2, 17q23 and 19q13. *Human Molecular Genetics.* 2016;25(13):2862-2872.
26. Ludwig KU, Böhmer AC, Bowes J, et al. Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip±cleft palate and cleft palate only. *Human Molecular Genetics.* 2017;26(4):829-842.
27. Ludwig KU, Ahmed ST, Böhmer AC, et al. 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 Genetics.* 2016;12(3):e1005914.
28. Yu Y, Zuo X, He M, et al. Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nature Communications.* 2017;8:14364.
29. Nikopensius T, Birnbaum S, Ludwig KU, et al. Susceptibility locus for non-syndromic cleft lip with or without cleft palate on chromosome 10q25 confers risk in Estonian patients. *European Journal of Oral Sciences.* 2010;118(3):317-319.
30. Beaty T, Taub M, Scott A, et al. Confirming genes influencing risk to cleft lip with/without cleft palate in a case–parent trio study. *Human Genetics.* 2013;132(7):771-781.
31. Leslie EJ, Taub MA, Liu H, et al. Identification of Functional Variants for Cleft Lip with or without Cleft Palate in or near PAX7, FGFR2, and NOG by Targeted Sequencing of GWAS Loci. *The American Journal of Human Genetics.* 2015;96(3):397-411.
32. Uslu VV, Petretich M, Ruf S, et al. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nature Genetics.* 2014;46(7):753-758.
33. Sharp GC, Stergiakouli E, Sandy J, Relton C. Epigenetics and orofacial clefts: a brief introduction. *The Cleft Palate-Craniofacial Journal.* 2017.
34. Juriloff DM, Harris MJ, Mager DL, Gagnier L. Epigenetic mechanism causes Wnt9b deficiency and nonsyndromic cleft lip and palate in the A/WySn mouse strain. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2014;100(10):772-788.

35. Plamondon JA, Harris MJ, Mager DL, Gagnier L, Juriloff DM. The clf2 gene has an epigenetic role in the multifactorial etiology of cleft lip and palate in the A/WySn mouse strain. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2011;91(8):716-727.
36. Alvizi L, Ke X, Brito LA, et al. Differential methylation is associated with non-syndromic cleft lip and palate and contributes to penetrance effects. *Scientific Reports.* 2017;7:2441.
37. Ouellette EM, Rosett HL, Rosman NP, Weiner L. Adverse effects on offspring of maternal alcohol abuse during pregnancy. *New England Journal of Medicine.* 1977;297(10):528-530.
38. Jones K, Smith D. Recognition of the fetal alcohol syndrome in early infancy. *The Lancet.* 1973;302(7836):999-1001.
39. Pitkin RM. Folate and neural tube defects. *The American Journal of Clinical Nutrition.* 2007;85(1):285S-288S.
40. Bailey LB, Stover PJ, McNulty H, et al. Biomarkers of nutrition for development—folate review. *The Journal of Nutrition.* 2015:jn206599.
41. MRC Vitamin Study Research Group. Prevention of neural tube defects: results of the Medical Research Council Vitamin Study. *The Lancet.* 1991;338(8760):131-137.
42. Hanrahan JP, Tager IB, Segal MR, et al. The effect of maternal smoking during pregnancy on early infant lung function. *American Review of Respiratory Disease.* 1992;145(5):1129-1135.
43. Harlap S, Davies AM. Infant admissions to hospital and maternal smoking. *The Lancet.* 1974;303(7857):529-532.
44. Weitzman M, Gortmaker S, Walker DK, Sobol A. Maternal smoking and childhood asthma. *Pediatrics.* 1990;85(4):505-511.
45. Thapar A, Fowler T, Rice F, et al. Maternal smoking during pregnancy and attention deficit hyperactivity disorder symptoms in offspring. *American Journal of Psychiatry.* 2003;160(11):1985-1989.
46. Kramer MS. Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization.* 1987;65(5):663.
47. Sebire NJ, Jolly M, Harris J, et al. Maternal obesity and pregnancy outcome: a study of 287 213 pregnancies in London. *International Journal of Obesity & Related Metabolic Disorders.* 2001;25(8).
48. Stothard KJ, Tennant PW, Bell R, Rankin J. Maternal overweight and obesity and the risk of congenital anomalies: a systematic review and meta-analysis. *JAMA.* 2009;301(6):636-650.
49. Watkins ML, Rasmussen SA, Honein MA, Botto LD, Moore CA. Maternal obesity and risk for birth defects. *Pediatrics.* 2003;111(Supplement 1):1152-1158.
50. Czeizel AE. The primary prevention of birth defects: multivitamins or folic acid? *International Journal of Medical Sciences.* 2004;1(1):50.
51. Shaw GM, Wasserman C, O'Malley C, Tolarova M, Lammer E. Risks of orofacial clefts in children born to women using multivitamins containing folic acid periconceptionally. *The Lancet.* 1995;346(8972):393-396.
52. Wilcox AJ, Lie RT, Solvoll K, et al. Folic acid supplements and risk of facial clefts: national population based case-control study. *BMJ.* 2007;334(7591):464.

53. Tolarova M, Harris J. Reduced recurrence of orofacial clefts after periconceptional supplementation with high-dose folic acid and multivitamins. *Teratology.* 1995;51(2):71-78.
54. Munger RG, Sauberlich HE, Corcoran C, Nepomuceno B, Daack-Hirsch S, Solon FS. Maternal vitamin B-6 and folate status and risk of oral cleft birth defects in the Philippines. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2004;70(7):464-471.
55. Little J, Gilmour M, Mossey P, et al. Folate and clefts of the lip and palate—a UK-based case-control study: Part I: Dietary and supplemental folate. *The Cleft Palate-Craniofacial Journal.* 2008;45(4):420-427.
56. Wehby G, Murray JC. Folic acid and orofacial clefts: a review of the evidence. *Oral Diseases.* 2010;16(1):11-19.
57. Johnson CY, Little J. Folate intake, markers of folate status and oral clefts: is the evidence converging? *International Journal of Epidemiology.* 2008;37(5):1041-1058.
58. Botto LD, Yang Q. 5, 10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: a HuGE review. *American Journal of Epidemiology.* 2000;151(9):862-877.
59. Shaw GM, Rozen R, Finnell RH, Todoroff K, Lammer EJ. Infant C677T mutation in MTHFR, maternal periconceptional vitamin use, and cleft lip. *American Journal of Medical Genetics Part A.* 1998;80(3):196-198.
60. Prescott N, Winter R, Malcolm S. Maternal MTHFR genotype contributes to the risk of non-syndromic cleft lip and palate. *Journal of Medical Genetics.* 2002;39(5):368-369.
61. Mostowska A, Hozyasz K, Jagodzinski P. Maternal MTR genotype contributes to the risk of non-syndromic cleft lip and palate in the Polish population. *Clinical Genetics.* 2006;69(6):512-517.
62. Zhu J, Ren A, Hao L, et al. Variable contribution of the MTHFR C677T polymorphism to non-syndromic cleft lip and palate risk in China. *American Journal of Medical Genetics Part A.* 2006;140(6):551-557.
63. Chevrier C, Perret C, Bahuau M, et al. Fetal and maternal MTHFR C677T genotype, maternal folate intake and the risk of nonsyndromic oral clefts. *American Journal of Medical Genetics Part A.* 2007;143(3):248-257.
64. Brandalize APC, Bandinelli E, Borba JB, Felix TM, Roisenberg I, Schüler-Faccini L. Polymorphisms in genes MTHFR, MTR and MTRR are not risk factors for cleft lip/palate in South Brazil. *Brazilian Journal of Medical and Biological Research.* 2007;40(6):787-791.
65. Jagomägi T, Nikopensius T, Krjutškov K, et al. MTHFR and MSX1 contribute to the risk of nonsyndromic cleft lip/palate. *European Journal of Oral Sciences.* 2010;118(3):213-220.
66. Boyles AL, Wilcox AJ, Taylor JA, et al. Oral facial clefts and gene polymorphisms in metabolism of folate/one-carbon and vitamin A: a pathway-wide association study. *Genetic Epidemiology.* 2009;33(3):247-255.
67. Kutbi H, Wehby GL, Uribe LMM, et al. Maternal underweight and obesity and risk of orofacial clefts in a large international consortium of population-based studies. *International Journal of Epidemiology.* 2016:dyw035.
68. Cedergren M, Källén B. Maternal obesity and the risk for orofacial clefts in the offspring. *The Cleft Palate-Craniofacial Journal.* 2005;42(4):367-371.

69. Blomberg MI, Källén B. Maternal obesity and morbid obesity: the risk for birth defects in the offspring. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2010;88(1):35-40.
70. Honein MA, Rasmussen SA, Reefhuis J, et al. Maternal smoking and environmental tobacco smoke exposure and the risk of orofacial clefts. *Epidemiology.* 2007;18(2):226-233.
71. Beaty TH, Maestri NE, Hetmanski JB, et al. Testing for interaction between maternal smoking and TGFA genotype among oral cleft cases born in Maryland 1992-1996. *The Cleft Palate-Craniofacial Journal.* 1997;34(5):447-454.
72. Shi M, Christensen K, Weinberg CR, et al. Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *The American Journal of Human Genetics.* 2007;80(1):76-90.
73. Zeiger JS, Beaty TH, Liang K-Y. Oral clefts, maternal smoking, and TGFA: a meta-analysis of gene-environment interaction. *The Cleft Palate-Craniofacial Journal.* 2005;42(1):58-63.
74. Lorente C, Cordier S, Goujard J, et al. Tobacco and alcohol use during pregnancy and risk of oral clefts. Occupational Exposure and Congenital Malformation Working Group. *American Journal of Public Health.* 2000;90(3):415.
75. Hackshaw A, Rodeck C, Boniface S. Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. *Human Reproduction Update.* 2011;17(5):589-604.
76. Källén K. Maternal smoking and orofacial clefts. *The Cleft Palate-Craniofacial Journal.* 1997;34(1):11-16.
77. Munger RG, Romitti PA, Daack-Hirsch S, Burns TL, Murray JC, Hanson J. Maternal alcohol use and risk of orofacial cleft birth defects. *Teratology.* 1996;54(1):27-33.
78. DeRoo LA, Wilcox AJ, Drevon CA, Lie RT. First-trimester maternal alcohol consumption and the risk of infant oral clefts in Norway: a population-based case-control study. *American Journal of Epidemiology.* 2008;168(6):638-646.
79. Shaw GM, Lammer EJ. Maternal periconceptional alcohol consumption and risk for orofacial clefts. *The Journal of Pediatrics.* 1999;134(3):298-303.
80. Werler MM, Lammer EJ, Rosenberg L, Mitchell AA. Maternal alcohol use in relation to selected birth defects. *American Journal of Epidemiology.* 1991;134(7):691-698.
81. Porter FD. Cholesterol precursors and facial clefting. *Journal of Clinical Investigation.* 2006;116(9):2322.
82. Engelking LJ, Evers BM, Richardson JA, Goldstein JL, Brown MS, Liang G. Severe facial clefting in Insig-deficient mouse embryos caused by sterol accumulation and reversed by lovastatin. *The Journal of Clinical Investigation.* 2006;116(9):2356-2365.
83. Munger RG, Tamura T, Johnston KE, Feldkamp ML, Pfister R, Carey JC. Plasma zinc concentrations of mothers and the risk of oral clefts in their children in Utah. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2009;85(2):151-155.
84. Tamura T, Munger RG, Corcoran C, Bacayao JY, Nepomuceno B, Solon F. Plasma zinc concentrations of mothers and the risk of nonsyndromic oral clefts in their children: A case-control study in the Philippines. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2005;73(9):612-616.
85. Krapels IP, Rooij IA, Wevers RA, et al. Myo-inositol, glucose and zinc status as risk factors for non-syndromic cleft lip with or without cleft palate in offspring: a case–

control study. *BJOG: An International Journal of Obstetrics & Gynaecology.* 2004;111(7):661-668.
86. Yoneda T, Pratt RM. Vitamin B6 reduces cortisone-induced cleft palate in the mouse. *Teratology.* 1982;26(3):255-258.
87. Davis SD, Nelson T, Shepard TH. Teratogenicity of vitamin B6 deficiency: omphalocele, skeletal and neural defects, and splenic hypoplasia. *Science.* 1970;169(3952):1329-1330.
88. Johansen AMW, Lie RT, Wilcox AJ, Andersen LF, Drevon CA. Maternal dietary intake of vitamin A and risk of orofacial clefts: a population-based case-control study in Norway. *American Journal of Epidemiology.* 2008;167(10):1164-1170.
89. Rodgers AB, Morgan CP, Leu NA, Bale TL. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. *Proceedings of the National Academy of Sciences.* 2015;112(44):13699-13704.
90. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. *The Lancet.* 1991;337(8746):867-872.
91. Wehby G, Cassell CH. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Diseases.* 2010;16(1):3-10.
92. Christensen BC, Houseman EA, Marsit CJ, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genetics.* 2009;5(8):e1000602.
93. Stanier P, Moore GE. Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts. *Human Molecular Genetics.* 2004;13(suppl 1):R73-R81.
94. Scapoli L, Palmieri A, Martinelli M, et al. Strong evidence of linkage disequilibrium between polymorphisms at the IRF6 locus and nonsyndromic cleft lip with or without cleft palate, in an Italian population. *The American Journal of Human Genetics.* 2005;76(1):180-183.
95. Shaffer JR, Orlova E, Lee MK, et al. Genome-wide association study reveals multiple loci influencing normal human facial morphology. *PLoS Genetics.* 2016;12(8):e1006149.
96. Marazita M. Subclinical features in non-syndromic cleft lip with or without cleft palate (CL/P): review of the evidence that subepithelial orbicularis oris muscle defects are part of an expanded phenotype for CL/P*. *Orthodontics & Craniofacial Research.* 2007;10(2):82-87.
97. Neiswanger K, Weinberg SM, Rogers CR, et al. Orbicularis oris muscle defects as an expanded phenotypic feature in nonsyndromic cleft lip with or without cleft palate. *American Journal of Medical Genetics Part A.* 2007;143(11):1143-1149.
98. Weinberg S, Naidoo S, Bardi K, et al. Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. *Orthodontics & Craniofacial Research.* 2009;12(4):271-281.
99. Zarate YA, Martin LJ, Hopkin RJ, Bender PL, Zhang X, Saal HM. Evaluation of growth in patients with isolated cleft lip and/or cleft palate. *Pediatrics.* 2010;125(3):e543-e549.
100. Wyszynski D, Sarkozi A, Vargha P, Czeizel A. Birth weight and gestational age of newborns with cleft lip with or without cleft palate and with isolated cleft palate. *Journal of Clinical Pediatric Dentistry.* 2004;27(2):185-190.

101. Marques IL, Nackashi JA, Borgo HC, et al. Longitudinal study of growth of children with unilateral cleft-lip palate from birth to two years of age. *The Cleft Palate-Craniofacial Journal.* 2009;46(6):603-609.
102. Menezes R, Vieira AR. Dental anomalies as part of the cleft spectrum. *The Cleft Palate-Craniofacial Journal.* 2008;45(4):414-419.
103. Ranta R. A review of tooth formation in children with cleft lip/palate. *American Journal of Orthodontics and Dentofacial Orthopedics.* 1986;90(1):11-18.
104. Reiser E, Skoog V, Gerdin B, Andlin-Sobocki A. Association between cleft size and crossbite in children with cleft palate and unilateral cleft lip and palate. *The Cleft Palate-Craniofacial Journal.* 2010;47(2):175-181.
105. Suzuki A, Takahama Y. Maxillary lateral incisor of subjects with cleft lip and/or palate: Part 1. *The Cleft Palate-Craniofacial journal.* 1992;29(4):376-379.
106. Suzuki A, Watanabe M, Nakano M, Takahama Y. Maxillary lateral incisors of subjects with cleft lip and/or palate: Part 2. *The Cleft Palate-Craniofacial Journal.* 1992;29(4):380-384.
107. Jamal GAA, Hazza'a AM, Rawashdeh MaA. Prevalence of dental anomalies in a population of cleft lip and palate patients. *The Cleft Palate-Craniofacial Journal.* 2010;47(4):413-420.
108. Nicholls W. Dental anomalies in children with cleft lip and palate in Western Australia. *European Journal of Dentistry.* 2016;10(2):254.
109. Akcam MO, Evirgen S, Uslu O, Memikoğlu UT. Dental anomalies in individuals with cleft lip and/or palate. *The European Journal of Orthodontics.* 2010;32(2):207-213.
110. Sharma RK, Nanda V. Problems of middle ear and hearing in cleft children. *Indian Journal of Plastic Surgery.* 2009;42(3):144.
111. Jocelyn LJ, Penko MA, Rode HL. Cognition, communication, and hearing in young children with cleft lip and palate and in control children: a longitudinal study. *Pediatrics.* 1996;97(4):529-534.
112. Sheahan P, Miller I, Sheahan JN, Earley MJ, Blayney AW. Incidence and outcome of middle ear disease in cleft lip and/or cleft palate. *International Journal of Pediatric Otorhinolaryngology.* 2003;67(7):785-793.
113. Skuladottir H, Sivertsen A, Assmus J, Remme AR, Dahlen M, Vindenes H. Hearing outcomes in patients with cleft lip/palate. *The Cleft Palate-Craniofacial Journal.* 2015;52(2):23-31.
114. Sell D, Grunwell P, Mildinhall S, et al. Cleft lip and palate care in the United Kingdom—the Clinical Standards Advisory Group (CSAG) Study. Part 3: speech outcomes. *The Cleft Palate-Craniofacial Journal.* 2001;38(1):30-37.
115. Hardin-Jones MA, Jones DL. Speech production of preschoolers with cleft palate. *The Cleft Palate-Craniofacial Journal.* 2005;42(1):7-13.
116. Lohmander A, Persson C. A longitudinal study of speech production in Swedish children with unilateral cleft lip and palate and two-stage palatal repair. *The Cleft Palate-Craniofacial Journal.* 2008;45(1):32-41.
117. Richman LC, McCoy TE, Conrad AL, Nopoulos PC. Neuropsychological, behavioral, and academic sequelae of cleft: early developmental, school age, and adolescent/young adult outcomes. *The Cleft Palate-Craniofacial Journal.* 2012;49(4):387-396.

118. Millard T, Richman LC. Different cleft conditions, facial appearance, and speech: relationship to psychological variables. *The Cleft Palate-Craniofacial Journal.* 2001;38(1):68-75.
119. Hunt O, Burden D, Hepper P, Johnston C. The psychosocial effects of cleft lip and palate: a systematic review. *European Journal of Orthodontics.* 2005;27(3):274-285.
120. Hunt O, Burden D, Hepper P, Stevenson M, Johnston C. Parent reports of the psychosocial functioning of children with cleft lip and/or palate. *The Cleft Palate-Craniofacial Journal.* 2007;44(3):304-311.
121. Wehby GL, Collet B, Barron S, Romitti PA, Ansley TN, Speltz M. Academic achievement of children and adolescents with oral clefts. *Pediatrics.* 2014:peds. 2013-3072.
122. Bille C, Winther JF, Bautz A, Murray JC, Olsen J, Christensen K. Cancer risk in persons with oral cleft—a population-based study of 8,093 cases. *American Journal of Epidemiology.* 2005;161(11):1047-1055.
123. Zhu JL, Basso O, Hasle H, Winther J, Olsen J, Olsen J. Do parents of children with congenital malformations have a higher cancer risk? A nationwide study in Denmark. *British Journal of Cancer.* 2002;87(5):524.
124. Steinwachs EF, Amos C, Johnston D, Mulliken J, Stal S, Hecht JT. Nonsyndromic cleft lip and palate is not associated with cancer or other birth defects. *American Journal of Medical Genetics Part A.* 2000;90(1):17-24.
125. Nishi M, Miyake H, Takeda T, Hatae Y. Congenital malformations and childhood cancer. *Pediatric Blood & Cancer.* 2000;34(4):250-254.
126. Zack M, Adami H-O, Ericson A. Maternal and perinatal risk factors for childhood leukemia. *Cancer Research.* 1991;51(14):3696-3701.
127. Dunkhase E, Ludwig KU, Knapp M, et al. Nonsyndromic cleft lip with or without cleft palate and cancer: Evaluation of a possible common genetic background through the analysis of GWAS data. *Genomics Data.* 2016;10:22-29.
128. Vogelaar IP, Figueiredo J, van Rooij IA, et al. Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. *Human Molecular Genetics.* 2012;22(5):919-926.
129. Machado RA, de Freitas EM, de Aquino SN, et al. Clinical relevance of breast and gastric cancer-associated polymorphisms as potential susceptibility markers for oral clefts in the Brazilian population. *BMC Medical Genetics.* 2017;18(1):39.
130. Bouvard V, Loomis D, Guyton KZ, et al. Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology.* 2015;16(16):1599-1600.
131. Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ.* 2004;328(7455):1519.
132. Mann C. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal.* 2003;20(1):54-60.
133. Tavani A, La Vecchia C. β-Carotene and risk of coronary heart disease. A review of observational and intervention studies. *Biomedicine & Pharmacotherapy.* 1999;53(9):409-416.
134. Lello L, Avery SG, Tellier L, Vazquez A, Campos Gdl, Hsu SD. Accurate Genomic Prediction Of Human Height. *arXiv preprint arXiv:170906489.* 2017.
135. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics.* 2017;101(1):5-22.

136. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics.* 2013;9(3):e1003348.
137. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology.* 2003;32(1):1-22.
138. Davey Smith G. Use of genetic markers and gene-diet interactions for interrogating population-level causal influences of diet on health. *Genes & Nutrition.* 2011;6(1):27.
139. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International Journal of Epidemiology.* 2016.
140. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Smith GD. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition.* 2016;103(4):965-978.
141. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine.* 2008;27(8):1133-1163.
142. Østergaard SD, Mukherjee S, Sharp SJ, et al. Associations between potentially modifiable risk factors and Alzheimer disease: a Mendelian randomization study. *PLoS Medicine.* 2015;12(6):e1001841.
143. Knowler WC, Williams R, Pettitt D, Steinberg AG. Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *American Journal of Human Genetics.* 1988;43(4):520.
144. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I 2 statistic. *International Journal of Epidemiology.* 2016;45(6):1961-1974.
145. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017;169(7):1177-1186.
146. Timpson NJ, Greenwood CM, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics.* 2017.
147. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology.* 2015;44(2):512-525.
148. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology.* 2016;40(4):304-314.
149. Hartwig FP, Smith GD, Bowden J. Robust inference in two-sample Mendelian randomisation via the zero modal pleiotropy assumption. *bioRxiv.* 2017:126102.
150. Robinson MR, Kleinman A, Graff M, et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour.* 2017;1:0016.
151. Rawlik K, Canela-Xandri O, Tenesa A. Indirect assortative mating for human disease and longevity. *bioRxiv.* 2017:185207.
152. Domingue BW, Fletcher J, Conley D, Boardman JD. Genetic and educational assortative mating among US adults. *Proceedings of the National Academy of Sciences.* 2014;111(22):7996-8000.

153. Waddington CH. Canalization of development and the inheritance of acquired characters. *Nature.* 1942;150(3811):563-565.
154. Ference BA, Yoo W, Alesh I, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *Journal of the American College of Cardiology.* 2012;60(25):2631-2639.
155. Holmes MV, Dale CE, Zuccolo L, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ.* 2014;349:g4164.
156. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics.* 1965;29(1):51-76.
157. Clarke T, Lupton M, Fernandez-Pujals A, et al. Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Molecular Psychiatry.* 2016;21(3):419-425.
158. Weiner DJ, Wigdor EM, Ripke S, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature Genetics.* 2017.
159. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics.* 2007;39(10):1181-1186.
160. International Consortium to Identify Genes and Interactions Controlling Oral Clefts 2010; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1.
161. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics.* 2007;81(3):559-575.
162. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nature Methods.* 2012;9(2):179-181.
163. Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.
164. Relton CL, Gaunt T, McArdle W, et al. Data resource profile: accessible resource for integrated epigenomic studies (aries). *International Journal of Epidemiology.* 2015;44(4):1181-1190.
165. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics.* 2012;4(3):325-341.
166. Allen NE, Sudlow C, Peakman T, Collins R. UK biobank data: come and get it. In: American Association for the Advancement of Science; 2014.
167. O'Connell J, Sharp K, Shrine N, et al. Haplotype estimation for biobank-scale data sets. *Nature Genetics.* 2016;48(7):817-820.
168. Consortium UK. The UK10K project identifies rare variants in health and disease. *Nature.* 2015;526(7571):82-90.
169. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics.* 2016;48(10):1279.
170. Bycroft C, Freeman C, Petkova D, et al. Genome-wide genetic data on~ 500,000 UK Biobank participants. *bioRxiv.* 2017:166298.

171. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology.* 2017.
172. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nature Genetics.* 2016;48(12):1544.
173. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics.* 2006;38(8):904-909.
174. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research.* 2009;19(9):1655-1664.
175. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods.* 2013;10(1):5-6.
176. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nature Genetics.* 2016;48(10):1284.
177. Paternoster L, Evans DM, Nohr EA, et al. Genome-wide population-based association study of extremely overweight young adults–the GOYA study. *PloS One.* 2011;6(9):e24303.
178. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13(1):86.
179. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS One.* 2012;7(7):e41361.
180. Stock NM, Humphries K, Pourcain BS, et al. Opportunities and challenges in establishing a cohort study: an example from cleft lip/palate research in the United Kingdom. *The Cleft Palate-Craniofacial Journal.* 2016;53(3):317-325.
181. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012;28(6):882-883.
182. Consortium G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660.
183. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics.* 2013;45(6):580-585.
184. Jansen R, Hottenga J-J, Nivard MG, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human Molecular Genetics.* 2017;26(8):1444-1451.
185. Weinberg SM, Raffensperger ZD, Kesterke MJ, et al. The 3D Facial Norms Database: Part 1. A web-based craniofacial anthropometric and image repository for the clinical and research community. *The Cleft Palate-Craniofacial Journal.* 2016;53(6):e185-e197.
186. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics.* 2008;9(4):255-266.
187. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics.* 2013;14(2):139-149.
188. Lander ES. The new genomics: global views of biology. *Science.* 1996;274(5287):536.
189. Reich DE, Lander ES. On the allelic spectrum of human disease. *TRENDS in Genetics.* 2001;17(9):502-510.

190. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *The American Journal of Human Genetics.* 2012;90(1):7-24.
191. Corder E, Saunders A, Strittmatter W, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science.* 1993;261(5123):921-923.
192. Eichner JE, Dunn ST, Perveen G, Thompson DM, Stewart KE, Stroehla BC. Apolipoprotein E polymorphism and cardiovascular disease: a HuGE review. *American Journal of Epidemiology.* 2002;155(6):487-495.
193. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development.* 2009;19(3):212-219.
194. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science.* 2007;316(5826):889-894.
195. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature.* 2015;518(7538):197-206.
196. Minster RL, Hawley NL, Su C-T, et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature Genetics.* 2016;48(9):1049.
197. Christensen K, Fogh-Andersen P. Cleft lip (±cleft palate) in Danish twins, 1970–1990. *American Journal of Medical Genetics Part A.* 1993;47(6):910-916.
198. Grosen D, Bille C, Petersen I, et al. Risk of oral clefts in twins. *Epidemiology* 2011;22(3):313.
199. Brito LA, Cruz LA, Rocha KM, et al. Genetic contribution for non-syndromic cleft lip with or without cleft palate (NS CL/P) in different regions of Brazil and implications for association studies. *American Journal of Medical Genetics Part A.* 2011;155(7):1581-1587.
200. Polderman TJ, Benyamin B, De Leeuw CA, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics.* 2015;47(7):702-709.
201. Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. *Nature Reviews Genetics.* 2002;3(11):872-882.
202. Carter C. Genetics of common disorders. In: *Problems of Birth Defects.* Springer; 1969:152-157.
203. Mitchell LE, Risch N. Mode of inheritance of nonsyndromic cleft lip with or without cleft palate: a reanalysis. *American Journal of Human Genetics.* 1992;51(2):323.
204. Fraser F. The multifactorial/threshold concept—uses and misuses. *Teratology.* 1976;14(3):267-280.
205. Fraser F. The genetics of cleft lip and cleft palate. *American Journal of Human Genetics.* 1970;22(3):336.
206. Carter C, Evans K, Coffey R, Roberts J, Buck A, Roberts MF. A three generation family study of cleft lip with or without cleft palate. *Journal of Medical Genetics.* 1982;19(4):246-261.
207. Melnick M, Bixler D, Fogh-Andersen P, Conneally PM, Elston RC. Cleft lip±cleft palate: an overview of the literature and an analysis of Danish cases born between 1941 and 1968. *American Journal of Medical Genetics Part A.* 1980;6(1):83-97.
208. Marazita ML, Goldstein AM, Smalley SL, Spence MA, Rao D. Cleft lip with or without cleft palate: Reanalysis of a three-generation family study from England. *Genetic Epidemiology.* 1986;3(5):335-342.

209. Chung C, Bixler D, Watanabe T, Koguchi H, Fogh-Andersen P. Segregation analysis of cleft lip with or without cleft palate: a comparison of Danish and Japanese data. *American Journal of Human Genetics.* 1986;39(5):603.
210. Hecht JT, Yang P, Michels VV, Buetow KH. Complex segregation analysis of nonsyndromic cleft lip and palate. *American Journal of Human Genetics.* 1991;49(3):674.
211. Carinci F, Pezzetti F, Scapoli L, et al. Nonsyndromic cleft lip and palate: evidence of linkage to a microsatellite marker on 6p23. *American Journal of Human Genetics.* 1995;56(1):337.
212. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* 1996;273(5281):1516-1517.
213. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease–common variant… or not? *Human Molecular Genetics.* 2002;11(20):2417-2423.
214. Cordell HJ, Clayton DG. Genetic association studies. *The Lancet.* 2005;366(9491):1121-1131.
215. Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. *American journal of human genetics.* 1996;59(5):983.
216. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *The American Journal of Human Genetics.* 1993;52(3):506.
217. Cardon LR, Bell JI. Association study designs for complex diseases. *Nature Reviews Genetics.* 2001;2(2):91-99.
218. Sun Y, Huang Y, Yin A, et al. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nature Communications.* 2015;6: 6414.
219. Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in~ 700,000 individuals of European ancestry. *bioRxiv.* 2018:274654.
220. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics.* 2011;88(1):76-82.
221. Loh P-R, Bhatia G, Gusev A, et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics.* 2015;47(12):1385.
222. Speed D, Cai N, Johnson MR, Nejentsev S, Balding DJ, Consortium U. Reevaluation of SNP heritability in complex human traits. *Nature genetics.* 2017.
223. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics.* 2012;91(6):1011-1021.
224. Palla L, Dudbridge F. A fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *The American Journal of Human Genetics.* 2015;97(2):250-259.
225. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics.* 2015;47(3):291-295.
226. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-753.

227. Mancuso N, Rohland N, Rand KA, et al. The contribution of rare variation to prostate cancer heritability. *Nature Genetics.* 2015;48(1):ng. 3446.
228. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences.* 2014;111(4):E455-E464.
229. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics.* 2010;11(6):446-450.
230. Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A. Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nature Genetics.* 2016.
231. Leslie EJ, Murray JC. Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clinical Genetics.* 2013;84(5):496-500.
232. Aylward A, Cai Y, Lee A, Blue E, Rabinowitz D, Haddad J. Using whole exome sequencing to identify candidate genes with rare variants in nonsyndromic cleft lip and palate. *Genetic Epidemiology.* 2016;40(5):432-441.
233. Hoebel A, Drichel D, van de Vorst M, et al. Candidate genes for nonsyndromic cleft palate detected by exome sequencing. *Journal of Dental Research.* 2017;96(11):1314-1321.
234. Kazeem G, Farrall M. Integrating case-control and TDT studies. *Annals of Human Genetics.* 2005;69(3):329-335.
235. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *The Lancet.* 2003;361(9357):598-604.
236. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics.* 2010;11(7):459-463.
237. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. *The American Journal of Human Genetics.* 2012;91(1):122-138.
238. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genetics.* 2009;5(10):e1000686.
239. Elhaik E, Greenspan E, Staats S, et al. The GenoChip: a new tool for genetic anthropology. *Genome Biology and Evolution.* 2013;5(5):1021-1031.
240. Elhaik E, Tatarinova T, Chebotarev D, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature Communications.* 2014;5.
241. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010;26(17):2190-2191.
242. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research.* 2011;46(3):399-424.
243. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature.* 2010;467(7311):52-58.
244. Sivertsen Å, Wilcox AJ, Skjærven R, et al. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *BMJ.* 2008;336(7641):432-434.

245. IPDTOC Working Group. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC). *The Cleft Palate-Craniofacial Journal.* 2011.
246. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics.* 2015;47(11):1228-1235.
247. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics.* 2015;47(11):1236-1241.
248. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics.* 2017;33(2):272-279.
249. Group IW. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts (IPDTOC). *The Cleft Palate-Craniofacial Journal.* 2011;48(1):66-81.
250. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics.* 2012;28(19):2540-2542.
251. Kumar SK, Feldman MW, Rehkopf DH, Tuljapurkar S. Limitations of GCTA as a solution to the missing heritability problem. *Proceedings of the National Academy of Sciences.* 2016;113(1):E61-E70.
252. Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences.* 2014;111(49):E5272-E5281.
253. Marazita ML, Lidral AC, Murray JC, et al. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Human Heredity.* 2009;68(3):151-170.
254. Leslie EJ, Carlson JC, Shaffer JR, et al. Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. *Human Genetics.* 2017;136(3):275-286.
255. Bird A. Perceptions of epigenetics. *Nature.* 2007;447(7143):396-398.
256. Gibney E, Nolan C. Epigenetics and gene expression. *Heredity.* 2010;105(1):4.
257. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics.* 2003;33:245-254.
258. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics.* 2016.
259. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International Journal of Epidemiology.* 2012;41(1):161-176.
260. Joubert BR, Felix JF, Yousefi P, et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *The American Journal of Human Genetics.* 2016;98(4):680-696.

261. Küpers LK, Xu X, Jankipersadsing SA, et al. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *International Journal of Epidemiology.* 2015;44(4):1224-1237.
262. Kriaucionis S, Bird A. DNA methylation and Rett syndrome. *Human Molecular Genetics.* 2003;12(suppl_2):R221-R227.
263. Das PM, Singal R. DNA methylation and cancer. *Journal of Clinical Oncology.* 2004;22(22):4632-4642.
264. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biology.* 2013;14(10):3156.
265. Birney E, Smith GD, Greally JM. Epigenome-wide association studies and the interpretation of disease-omics. *PLoS Genetics.* 2016;12(6):e1006105.
266. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology.* 2016;17(1):61.
267. Richardson TG, Haycock PC, Zheng J, et al. Systematic Mendelian randomization framework elucidates hundreds of genetic loci which may influence disease through changes in DNA methylation levels. *bioRxiv.* 2017:189076.
268. Richardson TG, Zheng J, Smith GD, et al. Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. *The American Journal of Human Genetics.* 2017;101(4):590-602.
269. genome.sph.umich.edu/wiki/LiftOver.
270. Golding P, Jones and the ALSPAC Study Team. ALSPAC–the avon longitudinal study of parents and children. *Paediatric and Perinatal Epidemiology.* 2001;15(1):74-87.
271. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology.* 2012:111-127.
272. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Research.* 2017;45(4):e22-e22.
273. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv.* 2016:078972.
274. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research.* 2015:0962280215597579.
275. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology.* 2013;37(7):658-665.
276. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine.* 2016;35(11):1880-1906.
277. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics.* 2015;31(21):3555-3557.
278. Chun S, Casparino A, Patsopoulos NA, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics.* 2017;49(4):600-605.

279. Hallonet M, Hollemann T, Pieler T, Gruss P. Vax1, a novel homeobox-containing gene, directs development of the basal forebrain and visual system. *Genes & Development.* 1999;13(23):3106-3114.
280. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research.* 2001;29(1):137-140.
281. Butali A, Suzuki S, Cooper ME, et al. Replication of genome wide association identified candidate genes confirm the role of common and rare variants in PAX7 and VAX1 in the etiology of nonsyndromic CL (P). *American Journal of Medical Genetics Part A.* 2013;161(5):965-972.
282. de Aquino SN, Messetti AC, Bagordakis E, et al. Polymorphisms in FGF12, VCL, CX43 and VAX1 in Brazilian patients with nonsyndromic cleft lip with or without cleft palate. *BMC Medical Genetics.* 2013;14(1):53.
283. Toriyama M, Shimada T, Kim KB, et al. Shootin1: A protein involved in the organization of an asymmetric signal for neuronal polarization. *The Journal of Cell Biology.* 2006;175(1):147-157.
284. Wang Y, Sun Y, Huang Y, et al. Validation of a genome-wide association study implied that SHTIN1 may involve in the pathogenesis of NSCL/P in Chinese population. *Scientific Reports.* 2016;6:38872.
285. Mostowska A, Hozyasz KK, Wojcicka K, Biedziak B, Jagodzinski PP. Polymorphic variants at 10q25. 3 and 17q22 loci and the risk of non-syndromic cleft lip and palate in the polish population. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2012;94(1):42-46.
286. Carlson JC, Taub MA, Feingold E, et al. Identifying Genetic Sources of Phenotypic Heterogeneity in Orofacial Clefts by Targeted Sequencing. *Birth Defects Research.* 2017:1030-1038.
287. Paternoster L, Zhurov AI, Toma AM, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *The American Journal of Human Genetics.* 2012;90(3):478-485.
288. Liu F, Van Der Lijn F, Schurmann C, et al. A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genetics.* 2012;8(9):e1002932.
289. Cole JB, Manyama M, Kimwaga E, et al. Genomewide association study of African children identifies association of SCHIP1 and PDE8A with facial size and shape. *PLoS Genetics.* 2016;12(8):e1006174.
290. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. *Nature Communications.* 2016;7.
291. Boehringer S, Van Der Lijn F, Liu F, et al. Genetic determination of human facial morphology: links between cleft-lips and normal variation. *European Journal of Human Genetics.* 2011;19(11):1192.
292. Peng S, Tan J, Hu S, et al. Detecting genetic association of common human facial morphological variation using high density 3D image registration. *PLoS Computational Biology.* 2013;9(12):e1003375.
293. Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics.* 1950;35(2):212.
294. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557.

295. Greco M, Del F, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine.* 2015;34(21):2926-2940.
296. Hemani G, Tilling K, Smith GD. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics.* 2017;13(11):e1007081.
297. Watanabe K, Taskesen E, Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications.* 2017;8(1):1826.
298. Aspinall A, Raj S, Jugessur A, Marazita M, Savarirayan R, Kilpatrick N. Expanding the cleft phenotype: the dental characteristics of unaffected parents of Australian children with non-syndromic cleft lip and palate. *International Journal of Paediatric Dentistry.* 2014;24(4):286-292.
299. Boncinelli E. Homeobox genes and disease. *Current Opinion in Genetics & Development.* 1997;7(3):331-337.
300. Holland PW, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. *BMC Biology.* 2007;5(1):47.
301. Setó-Salvia N, Stanier P. Genetics of cleft lip and/or cleft palate: association with other common anomalies. *European Journal of Medical Genetics.* 2014;57(8):381-393.
302. van den Boogaard M-JH, Dorland M, Beemer FA, van Amstel HKP. MSX1 mutation is associated with orofacial clefting and tooth agenesis in humans. *Nature Genetics.* 2000;24(4):342.
303. Shi B, Losee JE. The impact of cleft lip and palate repair on maxillofacial growth. *International Journal of Oral Science.* 2015;7(1):14-17.
304. Wren Y, Miller LL, Peters TJ, Emond A, Roulstone S. Prevalence and predictors of persistent speech sound disorder at eight years old: findings from a population cohort study. *Journal of Speech, Language, and Hearing Research.* 2016;59(4):647-673.
305. Wren YE, Roulstone SE, Miller LL. Distinguishing groups of children with persistent speech disorder: Findings from a prospective population study. *Logopedics Phoniatrics Vocology.* 2012;37(1):1-10.
306. Shriberg LD. Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech, Language, and Hearing Research.* 1993;36(1):105-140.
307. Harrison S, Lewis SJ, Hall AJ, et al. Association of SNPs in LCP1 and CTIF with hearing in 11 year old children: Findings from the Avon Longitudinal Study of Parents and Children (ALSPAC) birth cohort and the G-EAR consortium. *BMC Medical Genomics.* 2015;8(1):48.
308. Niskar AS, Kieszak SM, Holmes A, Esteban E, Rubin C, Brody DJ. Prevalence of hearing loss among children 6 to 19 years of age: the Third National Health and Nutrition Examination Survey. *JAMA.* 1998;279(14):1071-1075.
309. Nagarajan R, Savitha V, Subramaniyan B. Communication disorders in individuals with cleft lip and palate: An overview. *Indian Journal of Plastic Surgery.* 2010.
310. Honein MA, Kirby RS, Meyer RE, et al. The association between major birth defects and preterm birth. *Maternal and Child Health Journal.* 2009;13(2):164-175.

311. Jezewski P, Vieira A, Nishimura C, et al. Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. *Journal of Medical Genetics.* 2003;40(6):399-407.
312. Satokata I, Maas R. Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. *Nature Genetics.* 1994;6(4):348-356.
313. Carozza SE, Langlois PH, Miller EA, Canfield M. Are children with birth defects at higher risk of childhood cancers? *American Journal of Epidemiology.* 2012;175(12):1217-1224.
314. Bjørge T, Cnattingius S, Lie RT, Tretli S, Engeland A. Cancer risk in children with birth defects and in their families: a population based cohort study of 5.2 million children from Norway and Sweden. *Cancer Epidemiology and Prevention Biomarkers.* 2008;17(3):500-506.
315. Johnson KJ, Lee JM, Ahsan K, et al. Pediatric cancer risk in association with birth defects: A systematic review. *PloS One.* 2017;12(7):e0181246.
316. Fisher PG, Reynolds P, Von Behren J, Carmichael SL, Rasmussen SA, Shaw GM. Cancer in children with nonchromosomal birth defects. *The Journal of Pediatrics.* 2012;160(6):978-983.
317. Vieira AR, Khaliq S, Lace B. Risk of cancer in relatives of children born with isolated cleft lip and palate. *American Journal of Medical Genetics Part A.* 2012;158(6):1503-1504.
318. Deschler B, Lübbert M. Acute myeloid leukemia: epidemiology and etiology. *Cancer.* 2006;107(9):2099-2107.
319. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nature Reviews Genetics.* 2009;10(6):353.
320. Marin-Padilla M. Structural organization of the cerebral cortex (motor area) in human chromosomal aberrations. A Golgi study. I. D1 (13–15) trisomy, Patau syndrome. *Brain Research.* 1974;66(3):375-391.
321. Hozyasz KK, Mostowska A, Wójcicki P, et al. Nucleotide variants of the cancer predisposing gene CDH1 and the risk of non-syndromic cleft lip with or without cleft palate. *Familial Cancer.* 2014;13(3):415-421.
322. Pharoah PD, Guilford P, Caldas C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology.* 2001;121(6):1348-1353.
323. Mostowska A, Hozyasz KK, Wójcicki P, Lasota A, Dunin-Wilczyńska I, Jagodziński PP. Association of DVL2 and AXIN2 gene polymorphisms with cleft lip with or without cleft palate in a Polish population. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2012;94(11):943-950.
324. Letra A, Bjork B, Cooper M, et al. Association of AXIN2 with non-syndromic oral clefts in multiple populations. *Journal of Dental Research.* 2012;91(5):473-478.
325. Letra A, Menezes R, Granjeiro JM, Vieira AR. AXIN2 and CDH1 polymorphisms, tooth agenesis, and oral clefts. *Birth Defects Research Part A: Clinical and Molecular Teratology.* 2009;85(2):169-173.
326. Liu W, Dong X, Mai M, et al. Mutations in AXIN2 cause colorectal cancer with defective mismatch repair by activating β-catenin/TCF signalling. *Nature Genetics.* 2000;26(2):146.

327. Lammi L, Arte S, Somer M, et al. Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer. *The American Journal of Human Genetics.* 2004;74(5):1043-1050.
328. Lesseur C, Diergaarde B, Olshan AF, et al. Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. *Nature Genetics.* 2016;48(12):1544-1550.
329. Bierut LJ, Goate AM, Breslau N, et al. ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Molecular Psychiatry.* 2012;17(4):445.
330. Clarke T-K, Adams MJ, Davies G, et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N= 112 117). *Molecular Psychiatry.* 2017;22(10):1376.
331. Mitchell RE, Hemani G, Dudding T, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 1, 13/11/2017. 2017.
332. Konishi Y, Stegmüller J, Matsuda T, Bonni S, Bonni A. Cdh1-APC controls axonal growth and patterning in the mammalian brain. *Science.* 2004;303(5660):1026-1030.
333. Berx G, Becker KF, Höfler H, Van Roy F. Mutations of the human E-cadherin (CDH1) gene. *Human Mutation.* 1998;12(4):226-237.
334. Boyles AL, DeRoo LA, Lie RT, et al. Maternal alcohol consumption, alcohol metabolism genes, and the risk of oral clefts: a population-based case-control study in Norway, 1996–2001. *American Journal of Epidemiology.* 2010;172(8):924-931.
335. Carter C. The inheritance of congenital pyloric stenosis. *British Medical Bulletin.* 1961;17(3):251-253.
336. Leslie EJ, Liu H, Carlson JC, et al. A genome-wide association study of nonsyndromic cleft palate identifies an etiologic missense variant in GRHL3. *The American Journal of Human Genetics.* 2016;98(4):744-754.
337. Feenstra B, Geller F, Krogh C, et al. Common variants near MBNL1 and NKX2-5 are associated with infantile hypertrophic pyloric stenosis. *Nature genetics.* 2012;44(3):334.
338. Claes P, Roosenboom J, White JD, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics.* 2018:1.
339. Lawlor D, Richmond R, Warrington N, et al. Using Mendelian randomization to determine causal effects of maternal pregnancy (intrauterine) exposures on offspring outcomes: Sources of bias and methods for assessing them. *Wellcome Open Research.* 2017;2.
340. Evans DM, Zhu G, Dy V, et al. Genome-wide association study identifies loci affecting blood copper, selenium and zinc. *Human molecular genetics.* 2013;22(19):3998-4006.
341. Evans DM, Brion MJA, Paternoster L, et al. Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genetics.* 2013;9(10):e1003919.
342. Zhao Q, Wang J, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:180109652.* 2018.
343. Brion M-JA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. *International Journal of Epidemiology.* 2012;42(5):1497-1501.
344. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology.* 2016;45(6):1866-1886.

345. Freathy RM, Kazeem GR, Morris RW, et al. Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index. *International Journal of Epidemiology.* 2011:dyr077.
346. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology.* 2015;181(4):251-260.