# Family-based genetic association models

Miriam Gjerdevik

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2020

UNIVERSITY OF BERGEN

# Family-based genetic association models

Miriam Gjerdevik

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 15.06.2020

Year: 2020

Title: Family-based genetic association models

Name: Miriam Gjerdevik

Print: Skipnes Kommunikasjon / University of Bergen

# Scientific environment and funding

The work in this PhD thesis was carried out in the Research Group for Genetic Epidemiology at the Department of Global Public Health and Primary Care at the Faculty of Medicine, University of Bergen (UiB), Norway, with secondary affiliation to the Department of Clinical Science at the Faculty of Medicine, UiB. The PhD scholarship was funded by the Faculty of Medicine, UiB. During my PhD period, I have been a member of the Norwegian research school in bioinformatics, biostatistics and systems biology (NORBIS).

### Supervisors

Professor Håkon Kristian Gjessing
Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway, and Department of Global Public Health and Primary Care, Faculty of Medicine, UiB

Professor Rolv Terje Lie
Department of Global Public Health and Primary Care, Faculty of Medicine, UiB

Professor Pål Rasmus Njølstad
Department of Clinical Science, Faculty of Medicine, UiB

Professor Øystein Ariansen Haaland
Department of Global Public Health and Primary Care, Faculty of Medicine, UiB

### Secondary position

I have held a 20% secondary position at the Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. I have also been affiliated with the Centre for Fertility and Health at the Norwegian Institute of Public Health, Oslo, Norway.

### International exchange

I spent three months (September–December 2017) at the Institute of Genetic Medicine, Newcastle University, UK, visiting Professor Heather J. Cordell and her research group. My research stay was funded by NORBIS and UiB.

### Funding

This work was partly supported by the Research Council of Norway through Biobank Norway 2 (Grant 245464/F50) and through its Centres of Excellence funding scheme (Grant 262700), as well as the Bergen Medical Research Foundation (Grant 807191). The funding bodies played no role in the design of the studies, analyses, or interpretation of data, nor in writing the papers in this thesis.

# Acknowledgments

First and foremost, I would like to acknowledge the Department of Global Public Health and Primary Care (IGS) and the Faculty of Medicine for supporting a PhD project on method development. I truly believe that good research stems from collaboration across different disciplines, and IGS has provided me with a solid, diverse, and inspiring research environment.

Most importantly, I would like to thank my supervisors for excellent guidance and support! To my eminent main supervisor, Håkon Kristian Gjessing, thank you for always being so patient. Your knowledge (on everything!) is impressive, and it has been a real honor having you as my main supervisor. I hope I have not scared you away, and that you will continue as my main supervisor in a postdoc. To my co-supervisor, Rolv Terje Lie, thank you for all your support and for always seeing the bigger picture in research. Thank you for bringing me along when developing a new introductory course in statistics at the Faculty of Medicine. Teaching can actually be fun, and I have really enjoyed working with you! To Pål Rasmus Njølstad, thank you for being willing to supervise me. I hope there will be new opportunities for exploring the rich data on type 2 diabetes. Thanks also to my co-supervisor, Øystein Ariansen Haaland, for all your enthusiasm and support. Your door is always open, and you are always eager to help.

I would also like to thank my co-authors Julia Romanowska and Astanand Jugessur. Together with Håkon, Rolv Terje, and Øystein, we have formed our own amazing, little research group. You always show great interest in my work, and you always provide me with sound feedback and great ideas! Thanks also to my co-author Nikolai Olavi Czajkowski, for important input on my third paper.

During my PhD period, I spent three months at Newcastle University, visiting Professor Heather Cordell and her research group. I am grateful to Heather for welcoming me to her group, and for all her input, ideas, and great feedback. Special thoughts go to Richard Howey and Rachel Queen for inviting my husband and me to your homes and for making our stay such a success.

Most of my days as a PhD student have been spent at the old hospital for lepers at Kalfarveien 31, a great building where I have gathered many great memories. I wish to thank all members of the Research Group for Genetic Epidemiology, and especially Tone Bjørge, for many motivating and inspiring discussions. I would also like to thank my fellow PhD students and fun colleagues at IGS. Turid, you always make sure I have eaten well. A special thanks to Maria and Hilde Kristin, my office mates, for coping with me every day and for always keeping a good spirit. I am grateful that we are still able to keep in touch, even though we do not work together anymore.

I would also like to thank the Department of Mathematics, and my previous supervisor, Professor Ivar Heuch, for introducing me to the world of research. I would also like to thank Dean Per Bakke for always encouraging me and for all the possibilities you have provided me. Special thoughts go to my former colleagues at the Department of Thoracic Medicine, Haukeland University Hospital, for all the laughter and joy and a lifetime of friends.

My thanks to friends and family will be given in Norwegian.

Tiina. Siden vi møttes på Lungeavdelingen, har vi vært uadskillelige. Vi har delt oppturer og nedturer, og jeg setter uendelig stor pris på deg. Sammen med Linda, har vi vært en uslåelig trio som vet å finne gleden i de små og enkle ting.

Therese. Ingen dødtid på jobben er forgjeves, for da er det Therese-tid på telefon. Du er alltid deg selv, sier alltid det du mener, og du får meg alltid til å le, selv i de mest ekstreme situasjoner.

Mamma, pappa, Adrian og Negin. Det absolutt beste jeg vet er lørdagskveldene rundt kjøkkenbordet, med god mat og drikke, bondebridge eller domino. Vi har det alltid så gøy sammen! Det er trygt og godt å ha min egen heiagjeng som jeg vet stiller opp for meg, uansett.

Henningen min. Det sies at ekteskap krever hardt arbeid, men å være gift med deg er det enkleste i verden. Du lager middag til meg hver eneste dag, vasker, rydder og tilrettelegger for at jeg skal få lov til å gjøre akkurat det jeg vil, når jeg vil. Du skaper den perferkte balansen i hverdagen, oppmuntrer meg og får meg til å smile hver eneste dag. Du er uendelig trygg og god, og du er den absolutt beste jeg vet!

# Abbreviations

| | |
|---|---|
| A | Adenine |
| C | Cytosine |
| c-c | Case-control design |
| CLO | Cleft lip only |
| CL/P | Cleft lip with or without cleft palate |
| CLP | Cleft lip and palate |
| CpG | Cytosine-phosphate-guanine |
| CPG | Conditional on parental genotypes |
| CPO | Cleft palate only |
| DNA | Deoxyribonucleic acid |
| EM algorithm | Expectation-maximization algorithm |
| EMIM | Estimation of maternal, imprinting, and interaction effects using multinomial modelling |
| FBAT | Family-based association test |
| fc | Case-father dyad design |
| FDR | False discovery rate |
| FWER | Family-wise error rate |
| G | Guanine |
| GxE | Gene-environment interaction |
| GxMe | Interaction between a SNP allele and DNA methylation |
| GPC | Genetic Power Calculator |
| GWAS | Genome-wide association studies |
| $H_0$ | Null hypothesis |
| $H_1$ | Alternative hypothesis |

| | |
|---|---|
| HWE | Hardy-Weinberg equilibrium |
| LD | Linkage disequilibrium |
| LEM | Log-linear and event history analysis with missing data using the EM algorithm |
| MAF | Minor allele frequency |
| mc | Case-mother dyad design |
| mc-mc | Case-mother dyads with control-mother dyads |
| meQTL | Methylation quantitative trait locus |
| mfc | Case-parent triad design |
| mfc-c | Case-parent triads with control offspring |
| mfc-mc | Case-parent triads with control-mother dyads |
| mfc-mfc | Case-parent triads with control-parent triads |
| MFG | Maternal-fetal genotype |
| MoBa | The Norwegian Mother, Father and Child Cohort Study |
| OR | Odds ratio |
| PAE | Parental allelic exchangeability |
| PO-LRT | Parent-of-origin likelihood ratio test |
| PoO | Parent-of-origin |
| PoOxE | Parent-of-origin-environment interaction |
| PoOxMe | Interaction between a parent-of-origin effect and DNA methylation |
| RR | Relative risk |
| RRR | Relative risk ratio |
| SNP | Single-nucleotide polymorphism |
| T | Thymine |
| TAT | Transmission asymmetry test |
| TDT | Transmission disequilibrium test |
| TRIMM | Triad multi-marker |

# Abstract

The high heritability and recurrence rates observed for several complex diseases justify the search for genetic risk factors. However, despite decades of intense and extensive research, the underlying genetic basis of most complex traits has not been fully deciphered. This unexplained genetic etiology underscores the need to examine etiologic disease mechanisms other than simple genetic effects alone, such as the effect of maternal genes or the effect of parental origin. Additionally, since genome-wide association studies (GWAS) are commonly underpowered due to the large number of single-nucleotide polymorphisms being tested, poorly designed and inadequately powered studies that are unable to capture most of the genetic variants underlying a trait might also contribute to the unexplained genetic etiology.

Family-based study designs have been introduced specifically for studies of genetic risk factors. The main study unit is the case-parent triad design, which involves genotyping cases (affected offspring) and both their biological parents. However, a variety of other child-parent configurations and population-based study designs are also amenable to genetic association studies, including (but not limited to) cases in combination with unrelated controls, case-mother dyads, and case-parent triads in combination with unrelated controls or control-parent triads. Large clinical and population-based biobanks and national health registries have created unique opportunities for genetic, epidemiological, and clinical research worldwide. Nonetheless, there is currently a lack of flexible models that accommodate family structure in data. Models that incorporate non-standard genetic effects, such as maternal effects and parent-of-origin effects, are warranted. Moreover, joint models that integrate genetic, environmental, and epigenetic risk factors are needed to elucidate their combined effect on disease.

This thesis focuses on models for analyzing GWAS data for binary disease traits as well as methods for maximizing the statistical power of such studies, allowing for a broad range of child-parent configurations in the calculations. Using maximum likelihood estimation in a log-linear model, we developed new methodology to detect parent-of-origin-environment interactions, a possible mechanism contributing to disease susceptibility that has not yet been sufficiently explored. The approach has been implemented in our **R** package Haplin. In the Haplin framework, we also developed an extensive setup for power and sample size calculations, both through

analytic approximations and Monte Carlo simulations, which is essential not only in study planning but also in understanding and interpreting statistical findings. Within the power calculation module, we also implemented a relative efficiency calculator. Relative efficiency measures allow a more informative and general design comparison than straightforward and standard power analyses. We aimed to optimize the study design in genetic association studies given the constraints of available resources, i.e., maximize the statistical power using the least sample collection and genotyping cost.

# List of publications

**Paper I** Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. *Ann Hum Genet.* 2018;82:60-73.

**Paper II** Gjerdevik M, Jugessur A, Haaland ØA, Romanowska J, Lie RT, Cordell HJ, Gjessing HK. Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls. *BMC Bioinformatics.* 2019;20:165.

**Paper III** Gjerdevik M, Gjessing HK, Romanowska J, Haaland ØA, Jugessur A, Czajkowski NO and Lie RT. Design efficiency in genetic association studies. *Stat Med.* 2020; Epub ahead of print. DOI: 10.1002/sim.8476.

# Contents

# 1 Background

The high heritability and recurrence rates observed for many complex traits and disorders justify the search for genetic risk factors. Genome-wide association studies (GWAS) scan single-nucleotide polymorphisms (SNPs) across the genome to identify genetic variants that are more common in individuals with a particular trait than in those without this trait. GWAS have identified hundreds of genetic variants associated with complex traits and diseases and improved our understanding of their genetic architecture [1–3]. Still, despite decades of genetic research, the causes of many complex traits and diseases remain largely unknown. An example is oral clefts, in which the genetic variants identified thus far explain only a small fraction of the observed familial clustering and assumed genetic variance [4–7]. This unexplained genetic etiology highlights the need to investigate etiologic disease mechanisms beyond simple genetic effects alone, such as the effect of maternal genes, parent-of-origin (PoO) effects, and interactions between genetic effects and environmental exposures. Furthermore, the large number of SNPs being tested in a GWAS may result in a high number of false negative association findings [8], and a larger proportion of disease heritability and phenotypic variation might be identified with increased statistical power.

Familiar epidemiological designs such as population-based case-control or cohort designs can be used to search for genetic risk factors [9, 10]. However, specific to genetic association studies is the use of family-based designs, in which cases (affected offspring) and their biological parents are genotyped [11, 12]. The family-based designs represent a challenge to the available statistical and computational methods, and proper models that account for family structure in data are needed.

This thesis includes three papers, all of which involve methods for analyzing GWAS data or maximizing the scientific gain of such studies, allowing for the inclusion of family-based designs. In Paper I [13], we developed methods for analyzing parent-of-origin-environment interactions (PoOxE), a yet unexplored but plausible cause of complex diseases. There is a lack of software for genetic power calculations accommodating family structure in data, complicating the interpretation of genetic association findings. A comprehensive framework for power calculations was developed in Paper II [14]. The statistical power may be increased through careful deliberation of possible study designs. In Paper III [15], we aimed to compare and optimize study designs for genetic association studies by assessing the relative efficiency of alternative designs.

This background will give an introduction to genetic association studies and GWAS in particular, followed by definitions of genetic effects and etiological scenarios. Family-based study designs relevant to this thesis will be described. I will then define the concept of statistical power and emphasize why its consideration is essential in the design of efficient genetic association analyses and in the interpretation of statistical results. Furthermore, a brief introduction to some of the basic statistical tests that first incorporated family-based study designs into their analyses will be given. Lastly, I will present our **R** package Haplin, a statistical software for genetic association analysis of binary disease traits [16, 17]. Haplin forms the basis for this thesis and is the framework in which our new methods and software developments have been implemented.

## 1.1 Genetic markers and single-nucleotide polymorphisms

A genetic marker can be described as a variation of a gene or a deoxyribonucleic acid (DNA) sequence at a locus, i.e., a known position on the chromosome, that can be used to identify individuals or populations, or to study associations between genes and a disease known or believed to have a genetic background. In the human genome, SNPs are the most abundant form of variation, in which an appreciable frequency (e.g., more than 1%) of individuals in the population differ by a single nucleotide (adenine (A), cytosine (C), guanine (G) or thymine (T)) in a segment of the DNA [10, 18]. For example, at a locus, most individuals might have the sequence CCT, whereas some might have the sequence CAT instead. Since there is a possibility of either having the alternative C or A, the second position is considered a SNP (see Figure 1). Each of two or more variants of a gene at a locus is termed an allele [19]. In humans, almost all SNPs are diallelic [18], meaning that only two alternatives of the nucleotide can occur. Thus, C and A are the possible alleles for the diallelic SNP in this example. The less common allele is termed the minor allele, and the proportion (i.e., relative frequency) at which it occurs in a given population is termed the minor allele frequency (MAF) [19]. SNPs occur very frequently in the human genome and thus provide a dense marker spacing. They are therefore commonly used as genetic markers to unravel the genetic basis of inherited diseases.

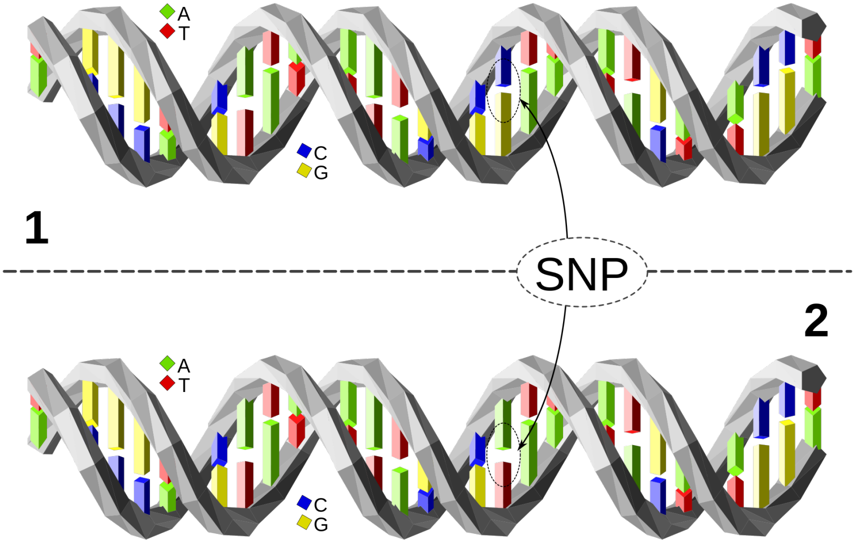

**Figure 1:** Illustration of a SNP. The two DNA molecules are different at a single base-pair location, where the upper DNA molecule has a C nucleotide and the lower has an A. SNP model by David Eccles (Gringer) [20]

Note, however, that several other types of genetic variation exist. For example, structural variants, including copy-number variants, translocations, or inversions of relatively large DNA segments, have been implicated for a number of diseases [2].

All individuals have two copies of each gene; one copy inherited from the mother and the other inherited from the father. Hence, for a SNP with alleles C and A, three genotypes are possible: CC, CA, and AA. In the simplest form of a genetic association analysis, the three genotypes can be used as exposure categories to investigate associations between genes and an inherited disease.

## 1.2 Mendelian and complex traits

Mendelian (monogenetic) traits are diseases or phenotypes caused by variation in a single gene, and the mode of inheritance can be dominant or recessive, autosomal, or linked to the X chromosome [21]. The alleles causing Mendelian traits are typically rare and highly penetrant, i.e., most individuals carrying the particular genetic variant also exhibit the associated disease (Figure 2). Mendelian traits are often recognized by their typical patterns of inheritance within families. Genetic linkage analysis, i.e., pedigree analysis of large families with multiple affected individuals, has therefore been successful in mapping the genetic basis of several Mendelian traits, such as Huntington's disease and cystic fibrosis [22, 23]. Two genetic loci on the same chromosome are linked if they are located near each other and thus tend to segregate together more often than what would be expected under independent inheritance. Hence, genetic linkage analysis quantifies the co-segregation of a marker locus and a trait locus among related subjects by studying within-family differences between markers and the trait in question [24].

Most traits are, however, not caused by variation in a single gene but have an architecture that is much more involved. Complex (multifactorial) traits are defined by the cumulative effect of multiple genes and possible interactions with environmental exposures and epigenetic factors [25, 26]. Examples of complex diseases are oral clefts, type 2 diabetes, Alzheimer's disease, and schizophrenia. A linkage analysis has low power to detect genes of moderate effect [27, 28]. Thus, although many complex traits are known to cluster in families, linkage studies have had limited success in mapping the multifactorial architecture underlying complex diseases.

## 1.3 Genetic association studies for complex traits

Genetic association studies are commonly used to identify SNPs (or other genetic variants) associated with complex traits. A marker allele is associated with a trait if the allele frequency is significantly higher or lower among affected individuals compared to what is expected from the general population (Figure 3) [24]. The candidate-gene approach to genetic association studies focuses on prespecified genes, based on a priori knowledge of its biological or statistical significance for the trait in question. However, the a priori knowledge is often limited, and candidate-gene studies have suffered from poor replication rates among reported significant associations [30–32]. In contrast, a GWAS scans the entire genome in thousands of individuals. Commonly used SNP arrays contain hundreds of thousands of SNPs [3], and a GWAS is therefore characterized as being a hypothesis-free approach. Nevertheless, the variants detected by a GWAS are mainly common alleles with low to moderate penetrance, i.e., only a small proportion of individuals with a given genotype exhibit its phenotypic effect (Figure 2). Typically, the identified alleles contribute to the

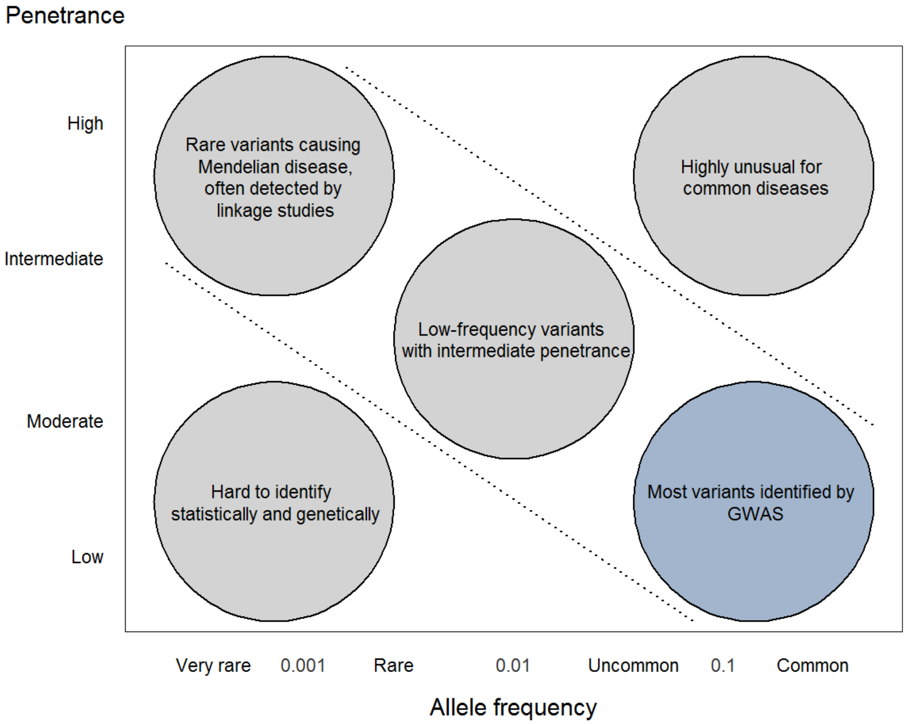
Penetrance
High
Intermediate
Moderate
Low
Rare variants causing Mendelian disease, often detected by linkage studies
Highly unusual for common diseases
Low-frequency variants with intermediate penetrance
Hard to identify statistically and genetically
Most variants identified by GWAS
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common
Allele frequency

**Figure 2:** Correlation between allele frequency, penetrance and disease susceptibility. An important aim of genetic research is to identify associations with the characteristics shown within the two diagonal lines. Most genetic variants identified by GWAS have low to moderate effect size and are relatively common in the population, as shown by the blue circle. Adapted from McCarthy et al. [29]

inherited component of complex diseases but cannot, even when combined, fully explain the total disease susceptibility [1, 29]. Low-frequency alleles with intermediate penetrance might explain more of the heritability, but increased sample sizes are needed to identify these variants through a GWAS [1, 33].

Genetic association methods are the main focus of this thesis and include both candidate-gene and genome-wide association analyses. The basic statistical analyses are the same for both approaches, i.e., each SNP is analyzed in the same manner regardless of being a candidate SNP or part of a genome-wide scan. The main statistical difference is, however, in post-processing and interpretation of results, especially due to multiple testing issues induced by a GWAS approach. Note that an association does not imply that the marker allele itself is the disease-causing allele. It is more likely that an allelic association is due to linkage disequilibrium (LD), a non-random association between alleles at different loci on a chromosome in a natural breeding population, occurring, for example, when the marker allele and the actual disease-causing allele are so close that they are transmitted together more often than would be expected by chance [18, 24]. Alternative explanations could also be spurious associations caused by population stratification or simply a Type I error (false positive) [24]. These concepts will be elaborated in later sections (1.6 and 1.7).

## 1.4 The complexity of isolated oral clefts

Oral clefts are the most common craniofacial birth defect, with a prevalence of 1 in 700 livebirths worldwide [34]. Oral clefts are broadly categorized as to whether they affect the lip only (CLO), the palate only (CPO), or both lip and palate (CLP). Although debated, CLO and CLP have traditionally been analyzed combined, forming the single group of cleft lip with or without cleft palate (CL/P) [35]. While oral clefts are frequently seen in association with other anomalies or as part of recognized syndromes, the isolated form, i.e., non-syndromic and occurring without other congenital defects, constitutes approximately 70% of CL/P cases and 55% of CPO cases [36, 37]. Among first degree relatives, Sivertsen et al. [4] reported relative recurrence risks of 32 for isolated CL/P and 56 for isolated CPO, suggesting a stronger genetic component for CPO than for CL/P. The major role for genes is also supported by twin studies, where heritability estimates as high as 91% for isolated CL/P and

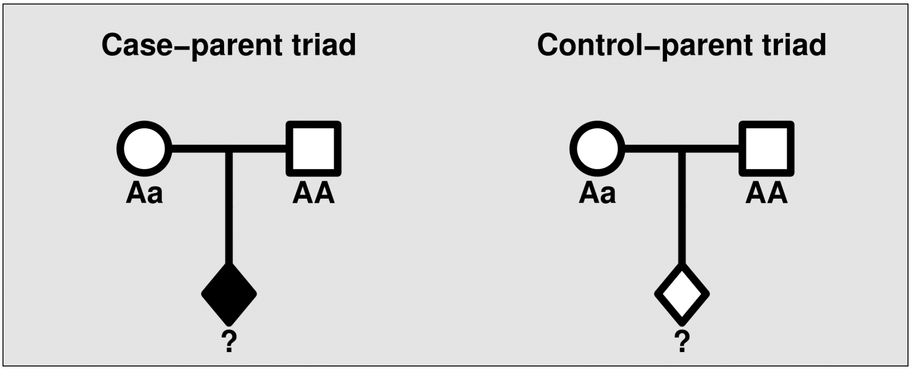

**Figure 3:** The hybrid design for family-based association analyses, consisting of affected offspring and their biological parents (case-parent triads) together with unaffected offspring and their biological parents (control-parent triads). The filled rhombus denotes the affected offspring. The probability of an $AA$ genotype is equal to that of $aA$ in both the case-parent triad and the control-parent triad, assuming Mendelian transmission. However, if there is an association between the genotype and the disease, the distribution among cases will differ from what would be expected under Mendelian transmission. The association approach tests for this asymmetry. Figure courtesy of Jugessur et al. [24]

90% for isolated CPO have been reported, with correspondingly small environmental factors (9% for isolated CL/P and 10% for isolated CPO) [6, 7]. Although the environmental contribution is likely to be smaller than the genetic component, the prevalence of oral clefts varies by ethnicity, geography, lifestyle, and environmental exposures [37, 38]. Thus, we cannot exclude the role of environmental risk factors and their possible interactions with genes. Moreover, because CL/P is more prevalent in males and CPO is more prevalent in females [37], it is reasonable to believe that also X-linked genes may contribute to the overall complexity of oral clefts.

## 1.5 Genetic effects and etiologic scenarios

To better understand disease biology, statistical methodologies that can differentiate between various casual models and disease mechanisms are needed. I will here introduce the genetic effects investigated throughout this thesis. The parameterization of penetrances is explained in Section 3.1.

### 1.5.1 Child effects

In the study of genetic effects, a relevant question relates to whether a variant allele inherited from one or both parents increases or decreases the risk of a disease, i.e., whether the genotype of an individual directly affects disease risk. This effect can be estimated from a case-control design, and terminology such as a “case genotype effect” has been used to describe this genetic effect in the literature [39]. However, the effect can also be fitted from the offspring in a case-parent triad. I will therefore refer to this as a “child effect” throughout this thesis, even though it is important to note that the offspring may be of any age, depending on the disease under study. In the study of pregnancy complications or birth defects (e.g., preeclampsia or isolated oral clefts), a child effect is sometimes referred to as a “fetal effect” [16, 40, 41]. This term was used in Paper I due to the application of new methodology to CPO data. In Paper III, we used the term “regular autosomal effect” to emphasize that the effect can also be estimated for late-onset diseases. Different modes of inheritance are possible for child effects, such as dominant, recessive, or multiplicative (log-additive) [42], as further described in Section 3.1. Although most association analyses have primarily targeted autosomal markers, the effect of offspring genes can also be linked to the X chromosome (Section 1.5.5). The terminology is somewhat confusing and ambiguous. However, the definition of child effects relates to the parameterization of penetrances, as described in Section 3.1.

### 1.5.2 Parent-of-origin effects

A PoO effect occurs if the phenotypic effect of a variant allele carried by an individual depends on its parent of origin. Hypothetically, an allele might be protective when derived from the mother but harmful when derived from the father. Because the effect of an allele in the child is modified by its parental origin, a PoO effect can be interpreted as a statistical interaction. This is in contrast to analyses of child effects, in which the two alleles in the child are considered to be functionally equivalent, i.e., the effect of a variant allele is assumed to be independent of whether it is inherited from the mother or the father. A PoO effect can be indicative of genomic imprinting, an epigenetic phenomenon where a variant allele carried by the child is expressed when inherited from one parent and silenced when inherited from the other [43–45]. Genomic imprinting may occur as an effect of different levels of DNA methylation (see Section 1.5.6) depending on parental origin, and it thus represents an exception to the classical Mendelian inheritance [46]. PoO effects have been implicated in numerous complex traits, e.g., attention deficit hyperactivity disorder [47] and type 2 diabetes [48, 49], suggesting that imprinted loci may explain parts of the phenotypic variation and disease heritability. However, few of these results have been validated by replication, and the impact of parent of origin has largely been overlooked [26, 48]. Studies of PoO effects are often insufficiently powered due to small sample sizes, and information on parental genotypes is not always available in data. Further research and better models to fit PoO effects are therefore requested [26].

### 1.5.3 Maternal effects

A maternal genetic effect occurs if a variant allele carried by the mother increases or decreases the risk of disease in her child, regardless of whether the allele is passed to her child [50]. This is distinct from our definitions of child and PoO effects, where we measure the effect of alleles carried by individuals themselves and not their mothers. The effect of maternal alleles may operate via mechanisms in the intrauterine environment, influencing the development of the fetus directly [51]. Maternal effects may therefore be particularly relevant for pregnancy conditions such as preeclampsia or birth defects that originate in fetal life [52]. However, conditions that depend on fetal development have also been associated with health outcomes throughout life [53, 54]. In recent years, the effect of maternal alleles has been estimated and discussed in a broader context too, and its relevance has been demonstrated even for an individual's educational attainment [55]. A maternal effect might be statistically confounded with a child or a PoO effect due to shared alleles between the mother and her child [39, 56]. Moreover, interaction effects might occur due to a maternal-fetal genotype incompatibility [57]. These underlying genetic mechanisms have different biological interpretations, and distinguishing between child, PoO, and maternal effects, as well as possible interactions, is important in advancing our understanding of the genetic architecture of complex traits [56, 57].

### 1.5.4 Gene-environment interactions

A genetic effect can be modified by an exposure or stratification variable such as an environmental risk factor, study site, and ethnicity. For example, maternal periconceptional vitamin use has been found to modify the association between SNPs and isolated CL/P [58]. This is referred to as a gene-environment interaction (GxE), and the genetic effect involved might be a child, a PoO, or a maternal effect. In the literature, the genetic effect most frequently referred to is a child effect. However, since epigenetic changes (e.g., DNA methylation, Section 1.5.6) can be modified by environmental factors, a search for interactions between PoO and environmental risk factors (PoOxE) might be particularly relevant [26]. We will use the abbreviation GxE without specific referral to the genetic effect in question, whereas PoOxE indicates that PoO is the genetic effect under scrutiny. A specific search for PoOxE has been the primary focus in several of our latest papers [59–61], and the methodology was developed in Paper I. The identification of GxE effects may not only improve our insights into the etiology of complex diseases but may also provide new opportunities to intervene on environmental risk factors alone, especially in population subgroups known to be genetically more susceptible to these exposure effects [60].

### 1.5.5 Effects of X-linked markers

Genes located on the X chromosome have distinctive patterns of inheritance since they are present in unequal numbers in males and females. A mother transmits one X chromosome to all of her children, whereas a father transmits his X chromosome to his daughters and his Y chromosome to his sons. The expression of X-linked markers is complex, and epigenetic processes such as DNA methylation (Section 1.5.6) may influence the dose effect in females. An example is X-inactivation, in which one of the two X chromosomes in females is silenced to ensure dosage compensation between the two sexes [62]. GWAS have mostly targeted autosomal markers, and analyses on the X-chromosome are underrepresented in the literature. This may be partly because most of the statistical methodology and software applied in genetic association studies were originally designed for the analysis of autosomal markers. However, since genetic variants on the X chromosome have been associated with several complex traits and diseases, methods and tools that accommodate the specific inheritance pattern of X-linked markers have been developed [63–67]. A search for genetic effects on the X chromosome is particularly relevant when a disease displays sex-specific differences in prevalence [67], as is seen for CL/P and CPO, systemic lupus erythematosus, and Sjögren's syndrome [37, 68]. Although most research on the X chromosome has been focusing on child effects, PoO and maternal effects may also be X-linked [69].

### 1.5.6 DNA methylation

DNA methylation is an epigenetic process where methyl groups are added to cytosine nucleotides, most commonly within cytosine-phosphate-guanine (CpG) dinucleotide motifs [61]. Although DNA methylation does not alter the underlying DNA sequence, it may still influence gene expression and manifest itself through various genetic effects such as PoO and X-inactivation. The methylation state is influenced by both environmental exposures and the DNA code itself. Nevertheless, the mechanisms through which gene-expression levels are affected are not yet fully understood [62, 70].

## 1.6 Study designs

A variety of family-based and population-based study designs are amenable to genetic association analyses. Relevant study designs include the standard case-control design, case-mother and case-father dyads, case-parent triads, and various case-family configurations in combination with unrelated controls or control families. Different study designs can accommodate different genetic effects, and each design has its own set of advantages and vulnerabilities. I will here give an introduction to the study designs relevant to this thesis.

### 1.6.1 The case-control design

Classic epidemiological designs such as the population-based case-control design (Figure 4a) are frequently used in genetic association analysis to identify child effects and their interactions with environmental or behavioral risk factors [71]. The allele frequencies of cases and controls are compared to detect variants associated with the disease under interrogation, and familiar statistical methods such as logistic regression or a chi-squared ($\chi^2$) test are commonly applied to test for effects [72]. However, population stratification might occur when cases and controls have been sampled from a heterogeneous population, where unrecognized subpopulations differ systematically in both allele frequencies and disease prevalence. Population stratification is a potential cause of false positive results in genetic association studies, but it could also mask a true association. Hence, additional control or correction for population stratification may be needed [73].

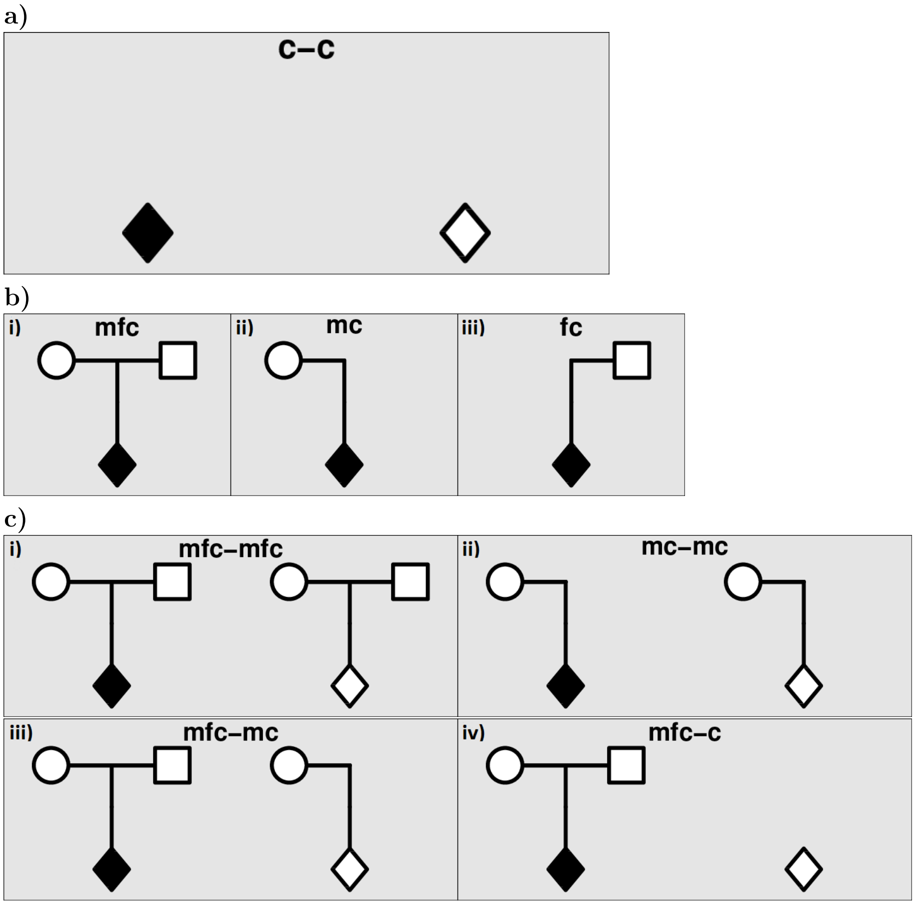

**Figure 4:** A selection of study designs for genetic association analyses. **a)** The case-control design (c-c); **b)** Various case-parent designs: i) Case-parent triad (mfc); ii) Case-mother dyad (mc); iii) Case-father dyad (fc); **c)** Various hybrid designs: i) Case-parent triad with independent control-parent triad (mfc-mfc); ii) Case-mother dyad with independent control-mother dyad (mc-mc); iii) Case-parent triad with independent control-mother dyad (mfc-mc); iv) Case-parent triad with independent control offspring (mfc-c)

### 1.6.2 The case-parent triad and dyad designs

In the late 1980s and early 1990s, Falk and Rubinstein [11] and Self et al. [74] observed that alleles associated with a given disease will occur more frequently in diseased offspring than what would be expected based on the parental allele distribution. Hence, parental genotypes of affected individuals could be used to study the association between genetic variants and a disease. The non-transmitted parental alleles would serve as individually matched genetic controls, i.e., so-called "pseudo-controls", thus eliminating the effects of population stratification. This insight gave rise to the family-based study designs [12, 75–78]. In the case-parent triad design, a sample of cases (affected offspring) and both their biological parents is genotyped. In the case-mother and case-father dyad designs, a sample of cases and their biological mothers or fathers is genotyped, respectively. The different designs are illustrated in Figure 4b. Besides removing bias due to population stratification, an inherent strength of the family-based designs is its ability to estimate PoO, or maternal effects from the information on parental genotypes. Whereas a child effect is estimated by comparing the allele frequencies of transmitted versus non-transmitted (pseudo-control) alleles, a PoO effect is primarily estimated in case families by comparing the frequencies of alleles transmitted from mother to child with the frequencies of alleles transmitted from father to child [14]. An allele working through the mother will be overrepresented in case-mothers compared with case-fathers [52]. Note that child, PoO and maternal effects can be estimated not only from case-parent triads but also from case-mother or case-father dyads. Nevertheless, there are also some drawbacks, and the family-based designs depend heavily on Mendelian transmission, which means that children are assumed to carry a random sample of the parental alleles. This fundamental Mendelian assumption must hold at the ages when children come under study. Moreover, unbiased estimates of maternal effects rely on "mating symmetry", i.e., we assume that the allele counts for mothers versus fathers are symmetric within parental mating types [52]. Another disadvantage of the family-based designs is the inability to estimate main effects of environmental exposures; interactions may be detected, but unrelated controls are required to determine whether the exposure is protective or detrimental [79]. Practical issues might also occur, such as obtaining DNA from parents if the disease is late onset. As a result, family-based designs may be genetically selective [80].

### 1.6.3 The hybrid design

To incorporate the advantages of the case-control and case-parent designs, Nagelkerke et al. [81] proposed a joint analysis of case-parent triads, unrelated cases and unrelated controls using generalized logistic (Poisson) regression. Their approach has been further explored and modified, and various other hybrid designs have been suggested [80, 82–86]. The full hybrid design comprises case-parent triads together with control-parent triads [85]. Weinberg and Umbach [80] and Vermeulen et al. [84] also use case-parent triads but propose different configurations of unrelated control families. Whereas the method by Weinberg and Umbach proposes genotyping parents of controls but not the controls themselves, Vermeulen et al. sample controls and their mothers. Since fathers may be hard to recruit, Shi et al. [83] proposed a case-mother/control-mother design. An overview of study designs and analysis features combining case-control and family data has been given by Infante-Rivard et al. [87]. Although the hybrid design combines the merits of both the case-control and case-parent triad designs, a straightforward combined analysis may be biased due to population stratification or non-Mendelian transmission, and corrections may be necessary to obtain valid estimates and inference. Different configurations of the hybrid design are illustrated in Figure 4c. Note that the hybrid designs do not necessarily involve the same number of case families as control families.

### 1.6.4 Notation

We have used the abbreviations in Figure 4 to denote the different study designs. The letters c, m, and f denote the child (case or control), mother and father, respectively. The left side of the hyphen denotes case families, whereas the right side denotes control families. For example, mfc denotes case-parent triads, mc denotes case-mother dyads, c-c denotes the case-control design, and mfc-mfc denotes the full hybrid design. We have used the term hybrid design to describe all constellations of study designs involving case families and unrelated control families, except for the c-c design.

## 1.7 Statistical power

A statistical hypothesis test is a method for drawing statistical inference from data in which statistical evidence for rejecting a hypothesis is summarized objectively.

In the classical (frequentist) approach to hypothesis testing [88], we formulate two competing hypotheses, a null hypothesis ($H_0$) and an alternative hypothesis ($H_1$), compute a test statistic using the observed data, and then decide whether to reject $H_0$ based on the calculated test statistic. The general formula for a test statistic can be written as

$$\text{Test statistic} = \frac{\text{Observed value - Hypothesized value}}{\text{Standard error of the observed value}}.$$

It is used to derive a $p$-value, defined as the probability of obtaining a difference at least as extreme as the one observed if $H_0$ is true, and we reject $H_0$ if the $p$-value is less than a preset threshold. Typically, $H_0$ refers to an effect size of zero (no difference), whereas a two-sided $H_1$ refers to a non-zero effect size.

When testing a null hypothesis, two types of errors can be made. The Type I error refers to falsely rejecting $H_0$, i.e., rejecting $H_0$ when it is true, and the probability of making a Type I error is defined as $\alpha$. The Type II error refers to the mistake of failing to reject $H_0$ when it is false. The probability of making a Type II error is defined as $1 - \gamma(\beta)$, where $\gamma(\beta)$ denotes the statistical power, and $\beta$ denotes the effect size. The statistical power is thus defined as the probability of correctly rejecting $H_0$ when $H_0$ is false and a true association exists. The definitions are summarized in Table 1.

**Table 1:** The two types of errors in hypothesis testing and their probabilities

| | | Decision: | |
|---|---|---|---|
| | | Do not reject $H_0$ | Reject $H_0$ |
| Truth: | $H_0$ is true | Correct decision<br>$1 - \alpha$ | Type I error<br>$\alpha$ |
| | $H_0$ is false | Type II error<br>$1 - \gamma(\beta)$ | Correct decision<br>$\gamma(\beta)$ |

The optimal study has small probabilities of making both types of errors. However, these probabilities are inversely related. The probability of making a Type I error, $\alpha$, is controlled by the researcher and is usually preset at the conventional threshold level of 0.05, known as the significance level of the test. Thus, the probability of making a Type II error, $1 - \gamma(\beta)$, and therefore also the statistical power, $\gamma(\beta)$, are subject to factors that cannot be controlled for, such as the true effect size or the MAF of a SNP. Nevertheless, measures can be taken to maximize the statistical power, e.g., increasing the sample size or optimizing the study design, although constraints of resources, such as money or the number of available cases, might limit these possibilities. In genetic association analyses, the effective sample size depends on the number of families, allele frequencies, and family design. These additional factors increase the complexity of power calculations.

### 1.7.1 Statistical power in a GWAS

The classical approach to hypothesis testing has been widely adopted in genetic association studies. Statistical power analyses are particularly important in a GWAS in order to maximize the scientific gains from the typically high genotyping and assay costs. They are also a prerequisite for optimal study design [8].

As previously mentioned, the conventional significance level of $\alpha = 0.05$ is commonly used to test a single null hypothesis. However, if $m$ independent hypothesis tests are performed, each at the $\alpha$ significance level, the probability of at least one false positive result is $1 - (1 - \alpha)^m \geq \alpha$ when the null hypothesis is true for all tests. If 1,000,000 tests are conducted, each at the 5% significance level, we expect 50,000 tests to be rejected by chance, even though no true association exists. The vast number of SNPs being tested in a GWAS leads to multiple testing issues, and a GWAS is therefore frequently underpowered. Moreover, most effect sizes reported from genetic association studies of complex traits are small, and empirical studies show that individual relative risks of disease are commonly below two [1, 89–91]. The small effect sizes further limit the power of a GWAS.

The statistical power provides valuable information when interpreting the results of a GWAS. Poor power may result in a large number of false negative findings, and a power analysis might shed light on non-significant associations by indicating whether the GWAS was inadequately powered. A power analysis may also indicate the smallest detectable effect size, given the sample size at hand [92]. Furthermore, poor power may increase the proportion of false positive findings among significant results. For example, in a study consisting of 1100 SNPs in which 100 have a true association with the disease, an expected number of 50 SNPs will be false positives at the 5% significance level, assuming no dependencies between the SNPs. The number of true positive findings is defined by $100 \cdot \gamma(\beta)$. That is, if $\gamma(\beta) = 1$, there are 100 true positive findings, which constitute 2/3 of the significant results (1/3 of the significant results are false positive findings). However, if $\gamma(\beta) = 0.5$, we expect 50 true positive findings, which constitute 1/2 of the significant results. The multiple testing burdens have resulted in the use of stringent significance thresholds in GWAS, and a genome-wide significance level of $5 \cdot 10^{-8}$ has been widely adopted to control the Type I error rate, thus allowing for multiple testing [8, 93]. Multiple testing issues will be further elaborated in Section 5.2.3.

### 1.7.2 An intuitive introduction to relative efficiency

As previously explained, a variety of child-parent configurations are amenable to genetic association studies. While different study designs can be compared directly by computing the power for a given set of parameter values, such calculations ignore the costs of data collection. For instance, a fixed number of complete case-parent triads could be compared with the same number of case-control pairs. Although the case-parent triad design requires 1.5 times the amount of genotyping relative to the case-control design (assuming the same number of cases and controls) [71], a straightforward power calculation would show identical power for the two alternatives. For example, using 500 case-parent triads, a relative risk (RR) of 1.3, and a MAF of 0.2 gives a power of 68% at the 5% nominal significance level. The same power is also obtained if we instead use 500 cases and 500 controls. Hence, a more informative and general design comparison can be achieved by studying the relative efficiency of two different study designs, defined as the ratio of sample sizes needed for each of the two designs to obtain the same significance level and power [94, Chapter 14]. This is equivalent to the ratio of variances of two separate parameter estimators, each estimator corresponding to one of the two study designs, taking into account the number of genotyped individuals within each design.

The concept of relative efficiency is closely related to that of statistical power and sample size. This relationship is illustrated in Table 2, in which we compared the efficiency of the full hybrid (mfc-mfc) design with that of the case-parent triad (mfc) design. For the mfc design, a design unit consists of one case child together with his/her biological parents (altogether three genotyped individuals). For the mfc-mfc design, we here used an equal number of case families and control families, and a design unit thus consists of one case-parent triad together with one control-parent triad (altogether six genotyped individuals). The total number of individuals required to obtain the desired power is calculated by multiplying the number of design units with the number of genotyped individuals within a unit. The relative efficiency is then computed by dividing the total number of individuals needed with

**Table 2:** The relationship between relative efficiency, statistical power, and sample size

| Power | mfc | | mfc-mfc | | Relative efficiency** |
|---|---|---|---|---|---|
| | Number of units | Number of individuals* | Number of units | Number of individuals* | |
| 0.6 | 415 | 1245 | 267 | 1602 | 0.78 |
| 0.7 | 523 | 1569 | 336 | 2016 | 0.78 |
| 0.8 | 665 | 1995 | 427 | 2562 | 0.78 |
| 0.9 | 890 | 2670 | 572 | 3432 | 0.78 |

The sample size is calculated for child effects using the Haplin function `snpSampleSize` with an RR of 1.3 and a MAF of 0.2 at the 0.05 nominal significance level. For the mfc-mfc design, we used an equal number of case families and control families

* The (total) number of individuals is computed by multiplying the number of design units with the number of genotyped individuals within each design unit (e.g., 415 case-parent triads consist of 1245 individuals)

** The relative efficiency is calculated by dividing the total number of individuals needed to obtain the desired power with the mfc design by that needed with the mfc-mfc design (e.g., 1245/1602=0.78). We see that the relative efficiency is constant across the different levels of power, favoring the mfc design the mfc design by that needed with the mfc-mfc design. It thus refers to a ratio of the number of genotyped individuals, not a ratio of the number of families or design units. We see that while 1995 individuals are needed for the mfc design to reach a power of 80%, 2562 individuals are required for the mfc-mfc design. The relative efficiency is 0.78, favoring the mfc design. In principle, the relative efficiency remains (close to) constant across the different levels of power and is therefore a useful measure for choosing the optimal study design. A more detailed discussion of relative efficiency is provided in Paper III, where we compared study designs asymptotically by using the concept of Pitman efficiency, i.e., by examining the variances obtained under the null hypothesis [95]. The Pitman efficiency is useful for preventing non-informative comparisons in situations where the effect size or sample size increases such that the power converges to 1.

## 1.8 Statistical methods for genetic association studies of binary disease traits

Genetic association studies have much in common with classic epidemiological studies of environmental risk factors. If the standard case-control design is used, the data can be analyzed in similar manners, for example, via standard $\chi^2$ tests for association or logistic regression [72]. Separate odds ratios can be estimated for the genotypes *aa* vs. *AA* and *aA* vs. *AA*, where lowercase indicates the minor allele. Alternatively, the genotypes can be grouped to assess dominant effects (*aa* and *aA* vs. *AA*), recessive effects (*aa* vs. *aA* and *AA*) or a dose-response relationship (e.g., coding *AA*, *aA*, and *aa* as 0, 1, and 2, respectively, and then applying a test for trend). With fewer parameters, such groupings would increase the statistical power, provided the model is correct.

Despite the similarities in analysis, several issues pertain specifically to genetic association studies. The family-based study designs have been proposed for genetic studies, and the transmission disequilibrium test (TDT) and related alternatives were introduced in the early 1990s to avoid spurious associations from population stratification [76]. In its simplest form, the TDT tests for over-transmission of an allele from heterozygous parents to affected offspring. It uses the standard McNemar test statistic for matched samples, given by

$$T = (n_{A,a} - n_{a,A})^2/(n_{A,a} + n_{a,A}),$$

where $n$ denotes the genotype counts as shown in Table 3. Under the null hypothesis of equal transmission, $T$ is asymptotically $\chi^2$ distributed with one degree of freedom. Only the off-diagonal elements of Table 3 are used in the calculations of $T$, and homozygous parents are therefore discarded. In its original form, $T$ cannot be calculated from families where the maternal or paternal genotype is missing [96], which potentially leads to a great loss of information. However, alternatives have been suggested to handle missing parental data, such as the 1-TDT [97].

An intuitive extension of the TDT for estimating PoO effects would be established by stratifying the frequencies of transmitted and non-transmitted alleles according to the parental origin. However, when accounting for parental origin, the ambiguous counts where both parents and offspring are heterozygous are often discarded. Moreover, there might be dependencies between parental transmissions from two heterozygous parents when the allele is associated with the disease [98], rendering the intuitive PoO approach statistically invalid when the model is not multiplicative. Although this can be avoided at the expense of power by discarding counts where both parents are heterozygous (the transmission asymmetry test (TAT) [99]), the TDT and its extensions are not able to separate the effects of alleles carried by the child, the mother, or both [50, 52].

To account for the drawbacks of the TDT-like approaches, flexible methods based on conditional logistic regression [100–102], log-linear [16, 52, 98, 99, 103], and multinomial modeling [104–106] have been proposed. As opposed to the TDT, which only calculates a single $p$-value, these models also produce relative risk estimates. A further advantage is the ease of generalization from the simplest situation of child effects to more advanced causal scenarios. For the assessment of PoO and maternal effects, a review and comparison of different statistical methodologies have been performed elsewhere [50]. The log-linear model of Gjessing and Lie [16] forms the basis of this thesis and will be described in greater detail in Section 3.1.

**Table 3:** Observed counts of transmitted and non-transmitted alleles for the TDT with data from affected offspring and both their parents

| | Non-transmitted allele | | |
|---|---|---|---|
| Transmitted allele | $A$ | $a$ | Total |
| $A$ | $n_{A,A}$ | $n_{A,a}$ | $n_{A,A} + n_{A,a}$ |
| $a$ | $n_{a,A}$ | $n_{a,a}$ | $n_{a,A} + n_{a,a}$ |
| Total | $n_{A,A} + n_{a,A}$ | $n_{A,a} + n_{a,a}$ | $2n$ |

The first index letter denotes the transmitted allele, and the second index letter denotes the non-transmitted allele. In total, there are $n$ offspring and $2n$ parents

## 1.9 The Haplin software

Several statistical tools for genetic association analysis exist that allow both estimation and testing of genotype relative risk parameters. A review of the most prominent programs is provided in Section 5.1, and I will here briefly introduce the Haplin software. Haplin provides the basis for this thesis into which all new methods and functionalities have been implemented. A detailed description of the underlying models is provided in several of our previous publications [13, 14, 16, 85] and will also be detailed in Section 3.

The **R** package [107] Haplin is based on log-linear modeling and provides a flexible framework for genetic association analyses of binary disease traits [16, 17]. A full maximum-likelihood model for estimation is implemented, and Haplin therefore provides explicit relative risk estimates with asymptotic standard errors and confidence intervals. Haplin enables the estimation of child effects, PoO effects, maternal effects, and GxE effects [13, 85]. Moreover, X chromosome analyses are easily performed, depending on the preassumed genetic model [66, 67, 69]. The basic log-linear model implemented in Haplin assumes Mendelian transmission, Hardy-Weinberg equilibrium (HWE), and random mating. Although the main unit of study is the case-parent triad, the log-linear model can be extended to include unrelated and unaffected controls or control families under the rare disease assumption [80]. Haplin uses the expectation-maximization (EM) algorithm [108] to account for unknown parental origin in ambiguous (uninformative) triads, e.g., when the mother, father, and child are all heterozygous for the same two alleles. The EM algorithm also accounts for missing parental genotypes, thus enabling analyses of case-mother or case-father dyads. The fundamental model in Haplin relates to a single multi-allelic locus. However, it can be adapted to the situation of multiple closely linked markers within a locus by statistically reconstructing haplotypes of unknown phase [16]. Furthermore, calculations can be performed in parallel, and Haplin is therefore well-suited for handling GWAS data. As part of this thesis, a complete setup for power, sample size, and relative efficiency calculations has recently been integrated into the log-linear framework and implemented as a new Haplin module. Installation details are given in Section 7 and on the Haplin website at `https://people.uib.no/gjessing/genetics/software/haplin`.

### 1.9.1 A Haplin example

An introduction to Haplin is most easily given for a child effect. We investigate a fictional SNP, here named rs123, with alleles $a$ and $A$, where $a$ is the less frequent. There are three possible genotypes: $AA$, $aA$ and, $aa$. We choose the more common genotype as our reference, $AA$, and estimate the relative risks $\mathrm{RR}_{aA}$ and $\mathrm{RR}_{aa}$ associated with the genotypes $aA$ and $aa$, respectively. If $a$ increases the risk, $\mathrm{RR}_{aA}$ and $\mathrm{RR}_{aa}$ should generally be larger than 1 (if the effect is recessive, $\mathrm{RR}_{aA} = 1$ and $\mathrm{RR}_{aa} > 1$ ). However, if $a$ decreases the risk, the estimates should generally be less than 1. We here assume a multiplicative dose-response model, i.e., $\mathrm{RR}_{aa} = \mathrm{RR}_{aA}^2$, although Haplin also allows estimation of both parameters separately.

The dataset `rs123_data` consists of 340 case-parent triads and 460 control-parent triads. A child effect is analyzed by the Haplin command

```
res <- haplin(rs123_data, response = "mult", design = "cc.triad",
       ccvar = 1, reference = "ref.cat").
```

The argument `response = "mult"` specifies a multiplicative dose-response relationship. The argument `design = "cc.triad"` specifies that our data contains a combination of case-parent triads and control-parent triads, and `ccvar = 1` points to the data column containing the case-control variable. The more frequent allele (genotype) is chosen as the reference category by the argument `reference = "ref.cat"`.

Haplin first outputs summary information on data and markers (here not shown), before continuing with the estimation results:

```
----Estimation results:----

Date of call: Mon Sep 09 09:37:41 2019

Number of triads: 800

Number of haplotypes: 2

Haplotype frequencies with 95% confidence intervals:
 Haplotype Frequency(%) lower upper
 A         90.19        88.98 91.29
 a          9.81         8.71 11.02
```

We see that the MAF is close to 10%. Haplin then outputs the relative risk estimates:

```
Single- and double dose effects (Relative Risk) with 95% confidence intervals:
Reference method: ref.cat
Reference category: 1 (Haplotype A)
Response model: mult

----Child haplotypes----
 Haplotype Dose     Relative Risk Lower CI   Upper CI   P-value
 A         Single   REF
 A         Double   REF

 a         Single   1.4           1.09       1.82       0.00918
 a         Double   1.97          1.18       3.32       0.00918
```

Relative to $A$ (or $AA$), carrying a single dose of $a$ increases the risk by 40%. Assuming a multiplicative dose-response model, carrying a double dose of $a$ gives a relative risk of $1.403^2 = 1.97$ (estimates with better precision are given by the command `haptable(res)`). The double-dose relative risk is not estimated freely, which is also demonstrated by the shared $p$-value. The result is significant at the 5% nominal level and is illustrated in Figure 5, obtained by the plotting function `plot(res)`. The fictional SNP, rs123, is simulated by the function `hapSim`, and the full code needed to obtain the data and run the analysis is given in Appendix I.

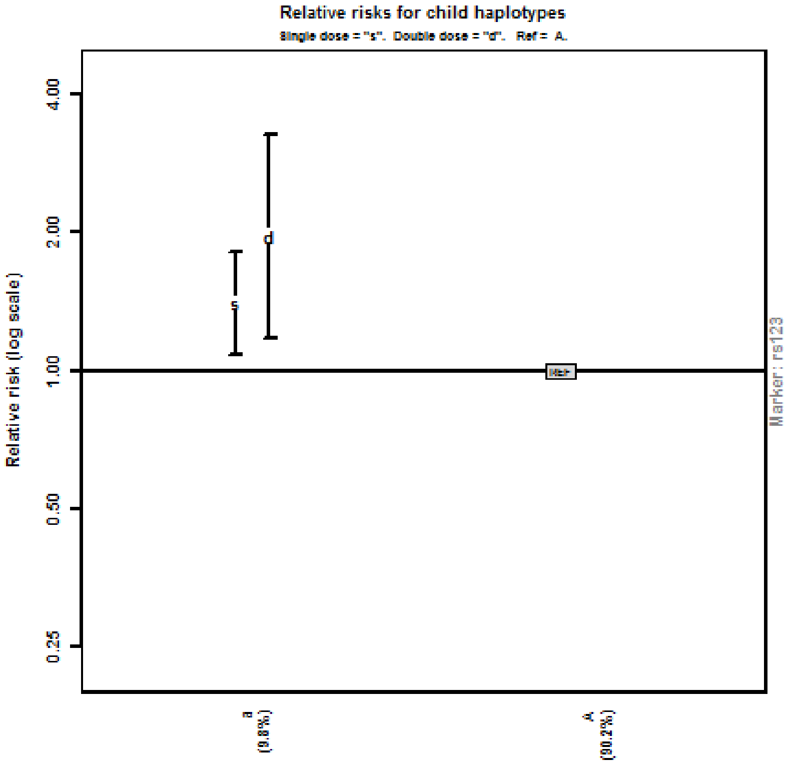

**Figure 5:** Estimated relative risks for child effects shown on a log-scale. Vertical bars represent 95% confidence intervals. Carrying either one or two copies of allele *a* increases the risk of disease, relative to the reference allele *A*

# 2 Objectives

Recent developments in genetic and epigenetic assays represent a great challenge to the available statistical and computational methods. In particular, important modeling challenges are:

- Appropriate models for family structure in data, in particular case children with parents (family data), with or without independent controls
- Models that integrate non-standard genetic effects beyond simple child effects, such as PoO and maternal effects
- Incorporation of genetic, environmental, and epigenetic risk factors in combined models that can elucidate their joint effect on disease
- Lack of framework for statistical power calculation based on the full triad design, including power calculations for child, PoO, and maternal effects, as well as interactions between genetic effects and environmental or epigenetic exposures

There is a general lack of implementation of such models, making it difficult to analyze GWAS data. Moreover, the lack of an extensive framework for statistical power analysis prevents optimal planning of study design and complicates the interpretation of statistical findings. In this context, the specific aims of the thesis are as follows.

- Develop and incorporate methods for assessing PoOxE effects in case-parent triads with or without unrelated controls (Paper I)
- Develop a framework for power and sample size analysis of genetic effects based on a variety of family-based study designs (Paper II)
- Provide insights into how relevant designs compare in terms of relative efficiency and optimize the study design for genetic association studies (Paper III)

The new methodologies and developments will be implemented in the Haplin framework, thus facilitating genetic association research of family-based data. The focus of this thesis is on binary disease traits. In Papers I—III, child, PoO, and maternal effects are primarily modeled assuming a multiplicative dose-response relationship (as outlined in Section 3.1), although other modes of inheritance can be fitted in the Haplin framework.

# 3 Statistical methods and material

This section will present statistical methods and material relevant to Papers I—III. The log-linear maximum likelihood approach forms the basis of Haplin, and a general introduction for child effects will be given. I will then outline how the model can be extended to handle PoO effects, maternal effects, and GxE effects for a locus with multiple alleles or haplotypes with unknown phase. I will also briefly explain how the EM algorithm can be applied to account for incomplete or missing data. Hypothesis testing in Haplin is mainly performed using a Wald test. I will introduce the Wald test statistic and explain how the statistical power of the test can be computed, both analytically and through Monte Carlo simulations. Next, I will describe the Haplin power functions, which have been written as part of this PhD project. In Paper II, an external validation of Haplin results was carried out by comparisons with the EMIM (Estimation of Maternal, Imprinting and interaction effects using Multinomial modelling) software. An introduction to EMIM is therefore given (for an overview of other statistical software for genetic association analysis, see Section 5.1). In Paper I and Paper III, data on CPO were used to illustrate the PoOxE test and relative efficiency measures, and details on the data material will be provided. I will then summarize the statistical methods and materials used for each paper and end this section with comments on ethical considerations.

## 3.1 The log-linear model

In this section, I will describe the underlying sampling and penetrance model of the log-linear likelihood approach. A more detailed derivation is provided in Gjessing and Lie [16].

We consider a single, multi-allelic locus with $K$ alleles $A_1$, $A_2$,..., $A_K$, with corresponding population allele frequencies $p_1$, $p_2$,..., $p_K$. The genotypes for the mother, father, and child are denoted by $M$, $F$, and $C$, respectively. Here, we assume that the child inherits the second allele from the mother and the second allele from the father. Thus, the full triad is denoted by $(M,F,C) = (A_iA_j,\ A_kA_l,\ A_jA_l) = (A_iA_j,\ A_kA_l)$. A case-parent triad is sampled through a case child, i.e., an affected offspring. Due to Bayes' theorem, the conditional probability of $(M,F,C)$ given disease in the child can be written as

$$P(M,F,C|D) = P(D|M,F,C)P(M,F,C)/P(D).$$

The disease prevalence $P(D)$ is unidentifiable owing to the sampling approach and functions only as a normalizing constant. The triad population frequency $P(M,F,C)$ can be expressed as $P(M,F,C) = P(C|M,F)P(M,F)$. The transmission probability, $P(C|M,F)$, depends only on Mendelian inheritance and is therefore trivial. The mating type probabilities, $P(M,F)$, are population quantities. Hence, if we also assume HWE and random mating, i.e., that allele and genotype frequencies will remain constant in a random-mating population, we have that $P(M,F,C) = p_i p_j p_k p_l$. The HWE restriction can be disturbed by factors such as population stratification, and the assumption can be avoided by considering the relative frequencies for the mating types [52, 99], or by including a multiplicative parameter that allows homozygotes to have a higher frequency in the population than what is expected under HWE [16]. Deviations from HWE and different parameterization models will be further discussed in Section 5.2.2.

The disease penetrance $P(D|M,F,C)$ is the probability of disease in the child conditional on the genotype of the case-parent triad. For child effects, we assume that the mating type $(M,F)$ is irrelevant when the genotype of the child is known. The penetrance can therefore be written as

$$P(D|M,F,C) = P(D|C) = P(D|A_j A_l) = B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl},$$

where $\mathrm{RR}_j$ and $\mathrm{RR}_l$ denote the relative risks associated with alleles $A_j$ and $A_l$, respectively, and where $B$ is the baseline risk level. Without loss of generality, we use $A_1$ as the reference allele and set $\mathrm{RR}_1 = 1$. Deviations from what would be expected from a multiplicative dose-response relationship is modeled by $\mathrm{RR}^*_{jl}$, where we set $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and $\mathrm{RR}^*_{jl} = 1$ if else. It follows that the full sampling model can be parameterized as

$$P(M,F,C|D) = p_i p_j p_k p_l \cdot B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl}/P(D). \tag{1}$$

For a diallelic SNP, the penetrance model is $P(D|A_1A_1) = B$, $P(D|A_1A_2) = B \cdot \mathrm{RR}$ and $P(D|A_2A_2) = B \cdot \mathrm{RR}^2\mathrm{RR}^*$. A recessive effect of $A_2$ would then be seen as $\mathrm{RR} = 1$ and $\mathrm{RR}^2\mathrm{RR}^* \neq 1$, and a dominant effect would mean that $\mathrm{RR} = \mathrm{RR}^2\mathrm{RR}^* \neq 1$. A multiplicative dose-response relationship would be seen as $\mathrm{RR}^2\mathrm{RR}^* = \mathrm{RR}^2$, i.e.,

$\text{RR}^*_j = 1$ for all $j$.

Let $n_{ijkl}$ denote the observed frequency of $(A_iA_j, A_kA_l)$. In the observed data, $n_{ijkl}$ relates to the triad probabilities $P(M, F, C|D)$. Thus, conditioning on disease in the child, the expected triad type frequencies can be written as

$$m_{ijkl} = E(n_{ijkl}) = \epsilon \cdot p_ip_jp_kp_l \cdot \text{RR}_j\text{RR}_l\text{RR}^*_{jl}, \qquad (2)$$

where $\epsilon$ is a normalizing constant. For a diallelic SNP, the theoretical multinomial distribution is shown in Table 4.

**Table 4:** Frequencies in case-parent triads for a diallelic SNP

| | Genotype | | | |
|---|---|---|---|---|
| Row number | Mother | Father | Child | Theoretical triad type frequency |
| 1 | $A_1\ A_1$ | $A_1\ A_1$ | $A_1\ A_1$ | $p_1^4$ |
| 2 | $A_2\ A_1$ | $A_1\ A_1$ | $A_1\ A_1$ | $p_1^3p_2$ |
| 3 | $A_1\ A_2$ | $A_1\ A_1$ | $A_2\ A_1$ | $\text{RR} \cdot p_1^3p_2$ |
| 4 | $A_2\ A_2$ | $A_1\ A_1$ | $A_2\ A_1$ | $\text{RR} \cdot p_1^2p_2^2$ |
| 5 | $A_1\ A_1$ | $A_2\ A_1$ | $A_1\ A_1$ | $p_1^3p_2$ |
| 6 | $A_2\ A_1$ | $A_2\ A_1$ | $A_1\ A_1$ | $p_1^2p_2^2$ |
| 7 | $A_1\ A_2$ | $A_2\ A_1$ | $A_2\ A_1$ | $\text{RR} \cdot p_1^2p_2^2$ |
| 8 | $A_2\ A_2$ | $A_2\ A_1$ | $A_2\ A_1$ | $\text{RR} \cdot p_1p_2^3$ |
| 9 | $A_1\ A_1$ | $A_1\ A_2$ | $A_1\ A_2$ | $\text{RR} \cdot p_1^3p_2$ |
| 10 | $A_2\ A_1$ | $A_1\ A_2$ | $A_1\ A_2$ | $\text{RR} \cdot p_1^2p_2^2$ |
| 11 | $A_1\ A_2$ | $A_1\ A_2$ | $A_2\ A_2$ | $\text{RR}^2 \cdot \text{RR}^* \cdot p_1^2p_2^2$ |
| 12 | $A_2\ A_2$ | $A_1\ A_2$ | $A_2\ A_2$ | $\text{RR}^2 \cdot \text{RR}^* \cdot p_1p_2^3$ |
| 13 | $A_1\ A_1$ | $A_2\ A_2$ | $A_1\ A_2$ | $\text{RR} \cdot p_1^2p_2^2$ |
| 14 | $A_2\ A_1$ | $A_2\ A_2$ | $A_1\ A_2$ | $\text{RR} \cdot p_1p_2^3$ |
| 15 | $A_1\ A_2$ | $A_2\ A_2$ | $A_2\ A_2$ | $\text{RR}^2 \cdot \text{RR}^* \cdot p_1p_2^3$ |
| 16 | $A_2\ A_2$ | $A_2\ A_2$ | $A_2\ A_2$ | $\text{RR}^2 \cdot \text{RR}^* \cdot p_2^4$ |

A normalizing constant must be included to ensure that the relative frequencies sum to 1. $p_1$ and $p_2$ are the allele frequencies corresponding to $A_1$ and $A_2$, respectively, i.e., $p_1 + p_2 = 1$. RR is the relative risk associated with $A_2$, using $A_1$ as the reference

Taking the logarithm, we have that

$$\log(m_{ijkl}) = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_3\boldsymbol{\beta}_3, \tag{3}$$

where

$$\boldsymbol{\beta}_1 = (\log(p_1), \dots, \log(p_K))^T,$$

$$\boldsymbol{\beta}_2 = (\log(\mathrm{RR}_2), \dots, \log(\mathrm{RR}_K))^T,$$

$$\boldsymbol{\beta}_3 = (\log(\mathrm{RR}_1^*), \dots, \log(\mathrm{RR}_K^*))^T,$$

and where $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are appropriate design matrices with dimensions $K^4 \times K$, $K^4 \times (K-1)$ and $K^4 \times K$, respectively. A more detailed explanation of the design matrix $\boldsymbol{X}$ is provided in Additional file 1 of Paper II, and an example will also be given in Section 3.5.2. Note that the intercept is not included in Eqn. (3) since the columns in $\boldsymbol{X}_1$ sum to a constant. The normalizing constant $\epsilon$ is accounted for when recovering the allele frequencies from $p_i = \exp(\beta_{1i})/\sum_j \exp(\beta_{1j})$, where the components of $\boldsymbol{\beta}_1$ have been estimated without the restriction $\sum_i p_i = 1$. We assume a Poisson distribution for the observed triad type frequencies $n_{ijkl}$, with expected cell values proportional to $P(M, F, C|D)$. The log-linear model in (3) can therefore be fitted by the use of standard software for Poisson regression (e.g., the `glm` function in **R** [107]), where the total log-likelihood is

$$l = \sum_{ijkl}(n_{ijkl}\log(m_{ijkl}) - m_{ijkl}).$$

Note that unrelated controls or control families are readily incorporated in the log-linear model. Let $\bar{D}$ denote the event that the child does not have the disease. Under the rare-disease assumption, we have that $P(\bar{D}|M, F, C) \approx 1$ and $P(\bar{D}) \approx 1$, resulting in $P(M, F, C|\bar{D}) \approx P(M, F, C)$. Hence, $P(M, F, C|\bar{D})$ can be expressed as the product of the population allele frequencies, $p_i p_j p_k p_l$. As illustrated in Figure 6, a relative risk is the measure of effect resulting from the case-parent triad design, whereas an odds ratio is the measure of effect resulting from the case-control design. In principle, the rare-disease assumption allows us to use relative risks and odds ratios interchangeably [85].

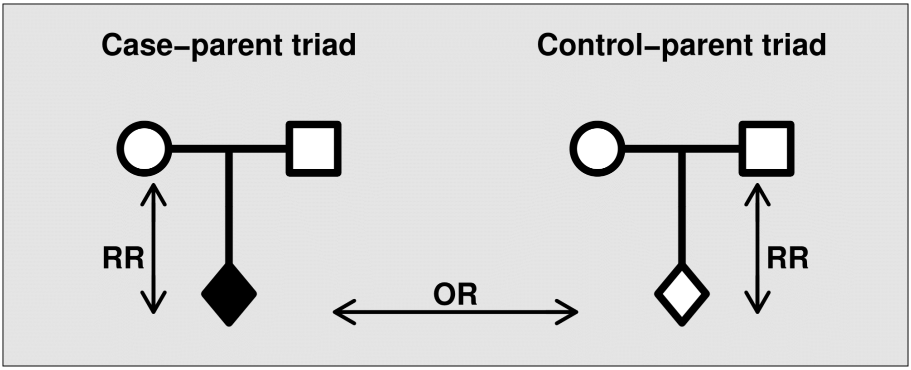

**Figure 6:** A relative risk (RR) is the effect measure resulting from the case-parent triad design, whereas an odds ratio (OR) is the effect measure resulting from the case-control design. Under the rare-disease assumption, the relative risks and odds ratios can be used interchangeably

### 3.1.1 Extensions to PoO and maternal effects

PoO and maternal effects are readily included in the log-linear model. For a PoO effect, we assign different effects to the alleles carried by the child, depending on the parent of origin. We then estimate the relative risk ratio $\mathrm{RRR}_j = \mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$, which measures the risk increase or decrease associated with allele $A_j$, relative to the reference allele, when the allele is transmitted from the mother as opposed to the father. To include the possible effects of maternal alleles, we assume that the alleles carried by the mother have a multiplicative effect in addition to the alleles in the child. The penetrance models are shown in Table 5.

To estimate PoO effects, we would need to know the parental origin, which is unknown for ambiguous triads, e.g., if all individuals in a triad are heterozygous for the same two alleles. To reconstruct parent of origin, our PoO model is therefore combined with the EM algorithm. The EM algorithm will be explained in Section 3.2.

### 3.1.2 Extensions to gene-environment interactions

With the additional assumption that, conditional on parental genotypes, the genotype of the child and the exposure status are independent, GxE effects can be analyzed using the case-parent triad design [79, 109]. We fit the log-linear model

**Table 5:** Parameterization of penetrances

| Effects | Parameterization of penetrances |
|---|---|
| Child | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl}$ |
| PoO | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl}$ |
| Child and maternal | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ |
| PoO and maternal | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ |

$B$ is the baseline risk level, typically associated with the reference allele; $\mathrm{RR}_j$ is the risk increase associated with allele $A_j$, relative to $B$; $\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ are the relative risks associated with allele $A_j$ when it is transmitted from the mother or from the father, respectively. We define a PoO effect as the relative risk ratio $\mathrm{RRR}_j = \mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$, which is a measure of the risk increase associated with $A_j$ when derived from the mother as opposed to the father; $\mathrm{RR}^*_{jl}$ measures deviations from what would be expected from a multiplicative model, i.e., $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and $\mathrm{RR}^*_{jl} = 1$ when $j \neq l$; $\mathrm{RR}^{(M)}_i$ is the relative risk associated with allele $A_i$ carried by the mother, and $\mathrm{RR}^{(M)*}_{ij}$ is the maternal double-dose parameter, interpreted analogously to $\mathrm{RR}^*_{ij}$. To ensure that the model is not overparameterized, we set $\mathrm{RR} = 1$ for the reference allele.
The table is adapted from Paper II and Paper III

separately in each exposure stratum for the genetic effect under study (i.e., child, PoO, or maternal effects) and apply a Wald-based post-test (described in Section 3.3) to examine whether the effect estimates deviate significantly across environmental strata. For child effects, the outcome of interest is the relative risk ratio, defined as $\text{RRR} = \text{RR}_{\text{exposed}}/\text{RR}_{\text{unexposed}}$. An $\text{RRR} > 1$ would mean that exposed children carrying the variant allele have an increased risk of disease relative to unexposed children carrying the variant allele. However, it is important to note that although interactions may be detected from the case-parent triad design, the main effect of an environmental exposure cannot be estimated without the addition of unrelated controls. A thorough derivation of the GxE and PoOxE models is the primary focus of Paper I and will, therefore, not be described in further detail herein. The conditional independence assumption underlying the GxE test can be relaxed when studying PoOxE effects. These constraints will be discussed in Section 5.2.1.

### 3.1.3 Haplotype estimation

A haplotype is defined within a region of a chromosome that is usually inherited as a single unit. It is a sequence of alleles from several closely linked SNPs or markers within a locus that tend to be inherited together. With the assumption of no recombination between the markers, the log-linear model for multiple alleles at a locus can be extended to a haplotype situation. If haplotype phase were known, the haplotype estimation would proceed as for a single, multi-allelic marker, treating each haplotype as a single allele. However, phase cannot be deduced for ambiguous triads. If one or several markers are ambiguous, we cannot, in general, deduce any of the haplotypes in the triad (with the exception of an individual being homozygous at all markers except for a single ambiguous marker) [16]. The number of ambiguous haplotypes will become substantial as the number of markers increases, and statistical reconstruction, for example via the EM algorithm, is necessary.

Haplin offers a sliding-window approach over a sequence of markers. This automates the analysis of a sequence of single SNPs, e.g., a GWAS analysis, or a haplotype analysis of a sequence of overlapping sliding windows. The rationale is that overlapping sliding windows may increase the chance of “bracketing” a causal variant if the haplotype has a SNP on each side of the variant. Nonetheless, loss of power is expected due to unknown haplotype phase and an increased number of degrees of freedom resulting from a larger number of alleles.

### 3.1.4 Analysis of X-linked markers

Several causal scenarios are relevant to an X-linked marker. In the log-linear setup of Haplin, various X-chromosome models may be fitted depending on the assumptions made about allele effects in females versus males, such as shared or different baseline risks, shared or different relative risks, and possible X-inactivation in females. Analyses of X-linked markers have not been of primary concern in Papers I—III and will, therefore, not be discussed any further in this thesis. For details on the possible parameterizations in Haplin, please consult our previous publications [66, 67, 69].

## 3.2 Using the EM algorithm to maximize the missing-data likelihood

In genetic association studies, incomplete information can originate from several sources. Genotype data could be missing due to failed genotyping, or family members, e.g., case-fathers or case-mothers, could be missing by design. Moreover, information is frequently lacking due to unknown phase or parental origin, such as when all three individuals in a case-parent triad are heterozygous (corresponding to rows 7 and 10 in Table 4). In the general population, one would actually expect this to occur in 12.5% of triads if both SNP alleles are equally likely (i.e., MAF = 0.5). Statistical methodology for handling unobserved variables or missing data is therefore essential in the analysis of genetic data.

Let $\boldsymbol{\beta}$ be the parameter vector and let $\boldsymbol{X}$ be an appropriate design matrix, as given in Eqn. (3). If we first assume that the full genotype of all triads can be observed, the number of each triad type, $\boldsymbol{n}$, e.g., corresponding to the rows of Table 4, can be described by independent Poisson distributions, where the expected number of triads in each row is given by $\boldsymbol{m} = \exp(\boldsymbol{X}\boldsymbol{\beta})$ (for a full explanation of the notation, formulas and dimensions, please consult Additional file 1 of Paper II). Hence, with complete information, $\boldsymbol{\beta}$ can be estimated by a straightforward maximization of the Poisson log-likelihood

$$l(\boldsymbol{\beta}) = \boldsymbol{n}^T \boldsymbol{X}\boldsymbol{\beta} - m_{\bullet},$$

where $m_{\bullet} = \boldsymbol{m}^T \mathbf{1}$.

With incomplete data, however, the likelihood contribution from a single observed (possibly ambiguous) triad $j$ is $\boldsymbol{a}_j^T\boldsymbol{p}$, where $\boldsymbol{p} = \boldsymbol{m}/m_{\boldsymbol{\cdot}}$ is the vector of cell probabilities in a multinomial model and $\boldsymbol{a}_j$ is defined as the ambiguity vector for the observed triad $j$, as explained in Additional file 1 of Paper II. If the mother, father, and child are all heterozygous for the same two alleles, $\boldsymbol{a}_j$ would be a vector with ones at positions 7 and 10, and zeros otherwise (Table 4). Thus, one would need to maximize the more difficult ambiguity log-likelihood

$$l_A(\boldsymbol{\beta}) = \sum_{j}^{N} (\log(\boldsymbol{a}_j^T\boldsymbol{m})) - m_{\boldsymbol{\cdot}}, \tag{4}$$

where $N = \boldsymbol{n}^T\mathbf{1}$ is the total number of observed triads. Usually, the maximum likelihood estimate of $\boldsymbol{\beta}$ has no closed form, and the ambiguity likelihood could be maximized directly via a search algorithm, such as an adapted Newton iterative approach [64]. However, Haplin uses instead the EM algorithm [108], which is a general and stable iterative optimization approach. It can be used to find maximum likelihood estimates with incomplete information, assuming that the missing genotypes are missing at random, i.e., independent of genotype. The EM algorithm is based on the idea of replacing the ambiguity log-likelihood $l_A(\boldsymbol{\beta})$ by a sequence of easier maximizations using the complete log-likelihood $l(\boldsymbol{\beta})$. The procedure is as follows. Starting from an (arbitrary) initial value of $\boldsymbol{\beta}$, we predict the expected number of triads in each row by the formula $\boldsymbol{m} = \exp(\boldsymbol{X\beta})$, pretending that the initial $\boldsymbol{\beta}$ is the true value. We then redistribute the ambiguity cells according to their expected values, given the observed total of ambiguity cells (e.g., the observed sum of rows 7 and 10 in the example above). This is the expectation (E) step, and the expectations can be calculated with both phase ambiguity and incomplete triads at the same time. In Haplin, the initial value of $\boldsymbol{\beta}$ is set to $\mathbf{0}$. Given the new distribution of cells, we then maximize $l(\boldsymbol{\beta})$ to find yet a new $\boldsymbol{\beta}$ estimate, using standard Poisson regression. This is the maximization (M) step. We continue the iteration process until the parameters converge. The ambiguity log-likelihood increases for every iteration, and the EM algorithm thus converges to the incomplete-data maximum likelihood estimate [108].

Even though the EM algorithm provides a valid estimate of $\boldsymbol{\beta}$, the variance-covariance matrix computed in each M step does not account for the extra uncertainty resulting from the incomplete data. The correct variance-covariance estimate, $\hat{\boldsymbol{\Sigma}}$, is the inverse of the observed Fisher information matrix, computed from $l_A(\hat{\boldsymbol{\beta}})$

in Eqn. (4). The derivation is given in Additional file 1 of Paper II. In Haplin, 50 EM iterations are used as default, which suffice to reach convergence in most situations. However, in PoO estimation based on genotype data from case-mother or case-father dyads, the number should be increased for small sample sizes (results not shown).

## 3.3 The Wald test

Our model is based on a maximum likelihood approach, and the Wald test can be used for hypothesis testing. Let $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_S]^T$ denote the combined vector of $S$ relative risk estimates on a log-scale with corresponding asymptotic variance-covariance estimate $\hat{\boldsymbol{\Sigma}}$. From standard asymptotic theory of log-linear models, we have that asymptotically, $\hat{\boldsymbol{\beta}}$ follows an approximate multivariate normal distribution, i.e.,

$$\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

as the sample size goes to infinity. The test of $r$ hypotheses on the $S$ parameters can be defined as a linear combination by using an $r \times S$ matrix $\boldsymbol{D}$. Asymptotically, we have that

$$\boldsymbol{D}\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{D}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{D}}),$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{D}} = \boldsymbol{D}\hat{\boldsymbol{\Sigma}}\boldsymbol{D}^T$. The Wald test statistic is then defined as

$$T = (\boldsymbol{D}\hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{D}}^{-1} (\boldsymbol{D}\hat{\boldsymbol{\beta}}).$$

Under the null hypothesis of $\boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{0}$, $T$ follows approximately a chi-squared distribution with $r$ degrees of freedom, $\chi^2(r)$. For the one-parameter case, the Wald statistic simplifies to

$$T = \frac{\hat{\beta}^2}{\hat{\sigma}^2} \sim \chi^2(1),$$

where $\hat{\sigma}^2$ is the asymptotic variance estimate of $\hat{\beta}$.

By defining an appropriate contrast matrix $\boldsymbol{D}$, a large number of hypotheses, including PoO, GxE, and PoOxE effects, can be tested. Note that for GxE and PoOxE effects, the strata are independent, and the variance-covariance matrix is therefore block diagonal. A detailed description is provided in Paper I. The Wald test is asymptotically equivalent to the likelihood ratio and score tests [110, Chapter 1]. However, the Wald test is calculated based on the parameter estimates and their variance-covariance matrix and is therefore computationally simpler in our situation.

## 3.4 Statistical power calculations

In Haplin, the statistical power can be computed based on the non-centrality parameter for the Wald test statistic or through Monte Carlo simulations. I will here give a general description of the two approaches.

### 3.4.1 Power of the Wald test

When the null hypothesis, $H_0$, is false, $T$ is asymptotically non-central chi-squared, $\chi^2(r, \lambda)$, with $r$ degrees of freedom and non-centrality parameter

$$\lambda = (\boldsymbol{D\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{D}}^{-1} (\boldsymbol{D\beta}), \tag{5}$$

i.e., the non-centrality parameter is calculated by replacing the estimated parameters in the Wald test statistic by their true values [111]. The statistical power of the Wald test can then be calculated analytically by the formula

$$\gamma = P(\chi^2(r, \lambda) > \chi^2_\alpha(r)), \tag{6}$$

where $\chi^2_\alpha(r)$ is the upper-$\alpha$ quantile of the chi-squared distribution with $r$ degrees of freedom. The power of rejecting $H_0$ at a fixed significance level $\alpha$ is thus completely determined by the number of degrees of freedom and the non-centrality parameter. For a fixed value of $r$, the power increases as the non-centrality parameter increases.

### 3.4.2 Monte Carlo simulations

The statistical power of a test can also be estimated by a Monte Carlo method in which the test procedure is replicated multiple times under given conditions. The observed power will then be the proportion of significant tests among the replicates. The following algorithm describes the Monte Carlo approach for power estimation, given a preselected value of the effect size and other necessary parameters concerning the test.

1. For each replicate, indexed by $j = 1, \ldots, J$:

   (a) Generate the $j^{\text{th}}$ random sample from the given distribution (under the conditions of the preset parameter values and effect size)

   (b) Calculate the test statistic from the $j^{\text{th}}$ sample

   (c) If $H_0$ is rejected at the nominal significance level $\alpha$, set $I_j = 1$. If else, set $I_j = 0$

2. Calculate the empirical power by computing the proportion of significant tests $\frac{1}{J}\sum_{j=1}^{J} I_j$

If the samples are generated from the null distribution, the proportion of significant tests will then be the observed Type I error rate, i.e., the attained significance level.

Although analytical power calculations are much more time-efficient than brute-force simulations, asymptotic results may not be valid for small to moderately sized datasets. The Monte Carlo simulation approach is, on the other hand, a completely general and robust statistical method for confirming software implementations, computing the empirical power and the empirical significance level, as well as for comparing statistical models and software. It is therefore a useful and valuable supplement to the asymptotic calculations.

## 3.5 Power and sample size analysis in Haplin

A considerable part of this thesis involves power analysis in genetic association studies. The theory underlying the calculations is not new. However, there has been a lack of software implementation, and a complete setup for power calculations based on different configurations of the hybrid design has not been available. I will here describe the Haplin framework for power analysis, which has been written as part of this PhD project.

### 3.5.1 `snpPower` and `snpSampleSize`

For single-SNP analyses of child effects, the statistical power and sample size can be computed using the Haplin functions `snpPower` and `snpSampleSize`. We assume a multiplicative dose-response relationship and count the number of "real" case alleles (the alleles transmitted from parents to affected offspring in case families), "real" control alleles (all alleles from the control families) and "pseudo-control" alleles (the alleles from case parents that have not been transmitted to the affected child).

The case (affected offspring) is the basis of the family-based designs and is always genotyped in our model. The total number of case alleles is then $N_1 = 2n_1$, where $n_1$ is the number of case families. The total number of control alleles can be written as $N_0 = a_1n_1 + a_0n_0$, where $a_1$ is the effective number of control alleles from a case family (i.e., pseudo-control alleles), $a_0$ is the effective number of control alleles from a control family, and $n_0$ is the number of control families.

A single case and a single control (without genotyping their parents) identify two case alleles and two control alleles, respectively. In this situation, $a_1 = 0$ and $a_0 = 2$. However, a complete case-parent triad has four alleles, two of which are transmitted to the case. The two non-transmitted alleles function as pseudo-controls, and a case-parent triad thus represents two case alleles and two control alleles ($a_1 = 2$). Moreover, the inclusion of a complete control-parent triad adds four control alleles ($a_0 = 4$).

The complexity of counting the alleles increases when case dyads or control dyads are included. If a case and only one of his/her parents are available, there are two case alleles and one control allele. However, we cannot always deduce which allele has been transmitted from the genotyped parent. Thus, $a_1 < 1$, and its value depends on the MAF and RR [112]. A similar argument applies if a control and only one parent are available for genotyping. We then have three control alleles but cannot deduce the parent of origin if both the control offspring and his/her parent are heterozygous. This reduces the effective number of control alleles ($a_0 < 3$). The results are summarized in Table 6.

When the total number of alleles is counted, the power calculations are similar to the approach used for a regular case-control design. `snpPower` calculates the power, $\gamma$, by using the asymptotic normal approximation for the natural logarithm of the odds ratio (OR), where the relative risks and odds ratios are used interchangeably due to the rare disease assumption. Hence, the formula is given by

$$\gamma = \Phi(z - z_{1-\alpha/2}) + \Phi(-z - z_{1-\alpha/2}),$$

where $\Phi(z) = P(Z \leq z)$, i.e., the standard normal distribution function, and

$$z = \ln(\mathrm{OR})/\sqrt{1/(N_1p_1(1-p_1)) + 1/(N_0p_0(1-p_0))}.$$

Here, $p_0$ is defined as the MAF within the control group, and $p_1 = p_0 \cdot \mathrm{OR}/(1 - p_0 + p_0 \cdot \mathrm{OR})$ is the MAF within the case group. Thus, for a given number of case families, control families, relative risks, minor allele frequencies, and Type I error rates, the power is readily computed.

The calculations are illustrated with a simple example. If we plan to include 100 case-parent triads in a study, both $N_1$ and $N_0$ equal 200. Assuming an RR (OR) of 2.0, a MAF ($p_0$) of 0.1 (here, the minor allele is the risk increasing allele), and a nominal significance level of 0.05, we have that $p_1 = 2/11$ and $z = \ln(2)/0.299 = 2.32$, resulting in a power of $\gamma = \Phi(2.32 - 1.96) + \Phi(-2.32 - 1.96) = 0.641$. The corresponding `snpPower` command and its output are

```
snpPower(cases = list(mfc = 100), controls = list(mfc = 0),
          RR = 2.0, MAF = 0.1)

  cases.mfc controls.mfc  RR MAF alpha    power
1       100            0   2 0.1  0.05 0.641071
```

Please consult the Haplin website or the `snpPower` help page in **R** for an explanation of the arguments and their options.

`snpSampleSize` is the inverse function of `snpPower`. For child effects, it calculates the number of case families and control families needed for a single SNP to obtain the desired power for specified family designs and given values of relative risks, minor allele frequencies, and Type I error rates. Applying the result of the above example, we obtain

```
snpSampleSize(fam.cases = "mfc", fam.controls = "no_controls",
               RR = 2.0, MAF = 0.1, power = 0.641071)

  fam.cases fam.controls RR MAF alpha    power case.families control.families
1       mfc  no_controls  2 0.1  0.05 0.641071           100                0
```

Explanations and documentation are given on the Haplin website and the help page in **R**.

Note that most of the functionality of `snpPower` is also covered by the more flexible Haplin function `hapPowerAsymp` (to be described next), which extends to power calculations of haplotype effects, PoO effects, maternal effects, GxE effects, etc. However, `snpPower` is somewhat easier to use and is therefore a valuable supplement for simple power calculations of single-SNP child effects.

**Table 6:** The effective number of control alleles for child effects in the single-SNP situation

| Genotyped individuals | Control family | Case family |
|---|---|---|
| | $a_0$ | $a_1$ |
| mfc | 4 | 2 |
| mf | 4 | - |
| mc or fc | $(3 - f(1))$* | $(1 - f(\mathrm{OR}))$* |
| m or f | 2 | - |
| c | 2 | 0 |

* The effective number of control alleles is derived by subtracting the subset of ambiguous dyads, where $f(\mathrm{OR}) = \frac{1 - \mathrm{MAF}}{\mathrm{MAF} \cdot \mathrm{OR}} / \left(1 + \frac{1 - \mathrm{MAF}}{\mathrm{MAF} \cdot \mathrm{OR}}\right)^2$, and where OR is set to 1 in control families

### 3.5.2 `hapPowerAsymp`

The asymptotic power of the Wald test is calculated by Eqn. (6), where the noncentrality parameter, $\lambda$, is given by Eqn. (5). The main difficulty is the calculation of $\boldsymbol{\Sigma}$, which is computed from the log-linear model accounting for transmission ambiguities and missing data. The derivation of $\boldsymbol{\Sigma}$ is given in Additional file 1 of Paper II.

The asymptotic power to detect an RR of 2.0, using 100 case-parent triads and a MAF of 0.1, is computed by the Haplin command

```
hapPowerAsymp(cases = c(mfc=100), haplo.freq = c(0.9,0.1), RR = c(1,2))
```

In order to calculate $\boldsymbol{\Sigma}$, we first need to compute the $\boldsymbol{\beta}$ values defined in Eqn. (3). We set $\beta_{11} = 0$ and calculate $\beta_{1j} = \log(p_j/p_1)$ for $j = 2, \ldots, K$. In this example, we have that $\boldsymbol{\beta}_1 = (0, -2.1972246)^T$ and $\boldsymbol{\beta}_2 = \beta_{21} = \log(2) = 0.6931472$ (the first allele is used as the reference, and $\boldsymbol{\beta}_2$ is therefore of length 1). Assuming a multiplicative dose-response effect, $\boldsymbol{\beta}_3$ is redundant. We then construct the 4-column matrix $\boldsymbol{G}$, which includes one column for each of the parental alleles, to list all possible triad genotypes. The matrix has dimensions $q \times 4$, where $q = l^4$ and $l$ is the number of alleles at a locus. For a diallelic SNP with alleles 1 and 2,

$$\boldsymbol{G} = \begin{array}{c} \begin{array}{cccc} A_{M1} & A_{M2} & A_{F1} & A_{F2} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 2 & 1 & 2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \end{pmatrix} \end{array},$$

with dimensions $16 \times 4$, where $A_M$ and $A_F$ denote the maternal and paternal alleles, ordered such that the second allele from each parent is transmitted to the child. From $\boldsymbol{G}$, we construct the corresponding $q \times p$ design matrix $\boldsymbol{X}$ for a log-linear model. It includes columns for estimating allele frequencies, child allele dose effects, maternal dose effects, etc., depending on the estimation model of interest. In our example, the design matrix is

$$\boldsymbol{X} = \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_c \end{array} \\ \begin{pmatrix} 4 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & 1 & 1 \\ 2 & 2 & 1 \\ 3 & 1 & 0 \\ 2 & 2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 3 & 1 \\ 1 & 3 & 2 \\ 0 & 4 & 2 \end{pmatrix} \end{array},$$

where the first and second columns count the number of alleles $A_1$ and $A_2$ in each row of $\boldsymbol{X}$, respectively, and the third column counts the number of variant alleles (here, $A_2$) inherited by the child. We then compute the expected number of triads in each row, $\boldsymbol{m} = \exp(\boldsymbol{X\beta})$, and the calculation of $\boldsymbol{\Sigma}$ now follows from Additional file 1 of Paper II (scaled to the correct sample size). Here, we are interested in the power to detect an RR of 2.0 for the minor allele $A_2$, which corresponds to the parameter $\beta_{21}$. Consequently, $\lambda = \beta_{21}^2/\sigma_{\beta_{21}}^2 = 0.6931472^2/0.08916667 = 5.388258$ and $P(\chi^2(1,\ 5.388258) > \chi^2_{0.05}(1)) = 0.641071$, which is equivalent to the power attained using `snpPower`.

Since `hapPowerAsymp` is integrated as part of the general analysis framework in Haplin, the value of $\mathbf{\Sigma}$ is readily attainable. This facilitates power calculations for different scenarios, such as for PoO and maternal effects, as well as extensions to GxE and haplotype effects. As Haplin is extended (e.g., to allow other study designs or genetic effects), corresponding power calculations will readily follow.

### 3.5.3 `hapRun` and `hapPower`

Haplin also includes a complete setup for power analysis through Monte Carlo simulations. The Monte Carlo approach to power calculation is described in Section 3.4.2, and the algorithm is implemented in Haplin via the two companion functions `hapRun` and `hapPower`. First, `hapRun` simulates haplotype data under the conditions of the given effect size and parameter values, in which triad genotypes are generated from the multinomial distribution (step 1a in the Monte Carlo algorithm). The multinomial probabilities are calculated by listing all possible genotype combinations in the triad format and then applying the sampling model in Eqn. (2) (or an adapted parameterization depending on the genetic effect in question). For a diallelic SNP, the multinomial distribution for a child effect is given in Table 4. If control-parent triads are added to the analysis, we set RR $= 1$ and RR$^* = 1$ in the multinomial. Next, `hapRun` runs Haplin on the simulated data, i.e., performs the statistical inference, and outputs the results (step 1b). This output is then fed to `hapPower`, which subsequently performs the power calculations by computing the proportion of $p$-values less than the nominal significance level (steps 1c and 2).

The power simulations in Haplin are general, and `hapRun` enables power calculations for a wider range of parameterization models than the current implementation of `hapPowerAsymp`. `hapRun` can also handle a composite of several child-parent configurations, and it permits missing individuals to be generated at random (see Additional file 2 of Paper II). Brute-force simulations are, however, time-consuming, and parallel processing has been implemented to speed the analyses. With four CPU cores (2.8 GHz per core), a regular power calculation of child effects takes less than 4 minutes using 10,000 data replicates of 200 case-parent triads for a diallelic SNP. A similar power calculation for PoO effects takes approximately 5 minutes. Examples and relevant Haplin commands for `hapRun` and `hapPower` are given in Additional file 2 of Paper II.

### 3.5.4 `hapRelEff`

`hapRelEff` computes the relative efficiency of two study designs. The variance for the relevant parameter estimator is calculated for each of the two designs and then compared. The number of genotyped individuals is taken into account. Thus, each individual is given the same cost regardless of disease status and regardless of being a child, a mother, or a father. Note that different costs (e.g., between cases and controls) can be inferred but are not considered in the current version of `hapRelEff`. The relative efficiency is calculated based on the asymptotic variance-covariance structure of the parameter estimator (Additional file 1 of Paper II), although a simulation procedure would be equally applicable.

I will illustrate the calculations with an example of PoO effects, comparing the mfc-mfc design (using an equal number of case and control families) with the mfc design, assuming a multiplicative dose-response relationship. For each design, we extract the relevant element from $\mathbf{\Sigma}$ (corresponding to the parameter of interest and scaled to a single design unit) and calculate

$$\frac{m_{\text{mfc}}\omega^2_{\text{mfc}}}{m_{\text{mfc-mfc}}\omega^2_{\text{mfc-mfc}}},$$

where $m_{\text{mfc}}$ and $m_{\text{mfc-mfc}}$ are the number of genotyped individuals for the mfc and mfc-mfc designs, respectively, and $\omega^2_{\text{mfc}}$ and $\omega^2_{\text{mfc-mfc}}$ are the variances representing a single design unit, as explained in Appendix 1 of Paper III. For this example, we observe that $\omega^2_{\text{mfc}} = \omega^2_{\text{mfc-mfc}}$, regardless of the MAFs and the relative risk values. Since $m_{\text{mfc}} = 3$ and $m_{\text{mfc-mfc}} = 6$, the relative efficiency equals 1/2, favoring the mfc design. Examples and relevant Haplin commands are given on the Haplin website and the R help page.

## 3.6 The EMIM software

For external validation of Haplin results, comparisons with other software are useful. Similar to Haplin, the companion programs PREMIM and EMIM are easy-to-use command-line tools for genetic association analysis of child, PoO, and maternal effects on autosomal markers, tailored to genotype data from several different child-parent configurations [104–106]. While PREMIM extracts genotype data from PLINK-format pedigree files (e.g., .ped or .bed files) [113] and generates the required input files for EMIM, EMIM performs the subsequent statistical analyses. PREMIM and EMIM are written in C++ and FORTRAN 77, respectively, and the run time is therefore faster than Haplin, which is implemented in **R**. The computational speed is an advantage for GWAS analysis. The statistical analyses in EMIM are performed using a multinomial modeling procedure which permits the simultaneous consideration of a variety of child-parent configurations. A range of different parameterization models and optional likelihood assumptions are allowed, including HWE and random mating. The equivalence between log-linear and multinomial models [114] implies that the modeling approaches used by Haplin and EMIM should provide equivalent inference. However, instead of fitting log-linear models for unobserved variables (ambiguities) via the EM algorithm, EMIM maximizes the multinomial likelihood directly via a maximization subroutine (MAXFUN, `http://darwin.cwru.edu/sage`). In order to perform the actual hypothesis tests, PREMIM and EMIM must be combined with external software (such as **R**), and there are no built-in commands for post-processing of results. Power computations are not implemented in EMIM, and several external steps are required to calculate the attained power of analysis, including data simulations and the computation of test statistics and corresponding *p*-values resulting from the EMIM analysis.

Information on EMIM and PREMIM is available from `https://www.staff.ncl.ac.uk/richard.howey/emim`. Details on the multinomial modeling approach and the various parameterization models used by EMIM have been described by Ainsworth et al. [104].

## 3.7 Case-parent triad study: illustration of analysis with cleft palate only data

From a previously published GWAS [115–117], genotype data and information on maternal periconceptional cigarette smoking were available for 550 isolated CPO families, including 466 complete case-parent triads. The CPO families stem from an international cleft collaboration which comprises both European/US and Asian populations. GWAS details have been provided in the original publication [115], and characteristics of the study population and information on quality control have been given by Haaland et al. [59]. The GWAS dataset is available from the dbGaP database (`https://www.ncbi.nlm.nih.gov/gap`) under study accession ID phs000094.v1.p1. We used the CPO data to demonstrate both our PoOxE test in Paper I and the relative efficiency estimates in Paper III. For the purpose of illustration, only a few SNPs were selected from the GWAS data, and a full genome-wide scan has not been performed as part of this thesis.

## 3.8 A brief overview of statistical methods and materials, Papers I—III

In all papers, Haplin is the main tool for analysis. Thus, log-linear models form the basis for the new methodological developments and software implementations. Statistical inference is based on the Wald test, and a multiplicative dose-response model has been assumed.

### 3.8.1 Paper I

We developed a new statistical and computational tool to estimate and test for PoOxE effects in a GWAS. The method can be described as a two-step approach. We first fit the log-linear model separately in each exposure stratum and then apply a Wald-based post-test to assess whether the PoO estimates deviate significantly across the exposure levels. The interaction approach was implemented in Haplin. As an illustration of the methodology, we applied the PoOxE test to top hits from previous published GWAS with case-parent triad data on CPO, assessing whether maternal smoking during the periconceptional period modifies the PoO effects. We used the same genetic triad data as applied in previous studies [59, 116, 117] and therefore stress that our examples and the corresponding results function only as an illustration and not as an independent replication of findings. Lastly, we evaluated the performance of the PoOxE test. Power calculations were mainly performed using asymptotic approximations (`hapPowerAsymp`). However, the attained significance level and the small-sample behavior were investigated through Monte Carlo simulations (`hapRun` and `hapPower`).

### 3.8.2 Paper II

We developed an extensive setup for power calculations in Haplin, including both analytical calculations using the non-centrality parameter for the Wald test statistic (Section 3.4.1 and 3.5.2) and a straightforward Monte Carlo simulation approach (Section 3.4.2 and 3.5.3). In Paper II, we compared the asymptotic power approximations (using `hapPowerAsymp`) to the power of analysis attained in simulations with Haplin (using `hapRun` and `hapPower`). For external validation, we further compared the results to the power of analysis attained in simulations using the EMIM software. For power analysis in EMIM, the Haplin function `hapSim` was used to simulate the genotype data. We then converted the data to standard PLINK-format files, which we subsequently fed into PREMIM and EMIM for analysis. Our primary focus was on child, PoO, and maternal effects.

### 3.8.3 Paper III

We provided insights into how relevant study designs compare in terms of relative (Pitman) efficiency and illustrated the methodology with extensive analyses for a range of genetic effects and etiologic scenarios based on asymptotic approximations. Our main focus was on child (regular autosomal), PoO, and maternal effects, and both single SNPs and haplotypes were assessed. Moreover, to facilitate relative efficiency analyses in other scenarios, we implemented the calculations as an easy-to-use function in Haplin (`hapRelEff`, Section 3.5.4). As a demonstration, we also compared the empirical efficiency of the case-mother dyad design with that of the case-parent triad design using preselected SNPs from the CPO data.

## 3.9 Ethical considerations and consents

**Paper I and Paper III** Ethics approvals for the cleft consortium were obtained from the respective ethics committees at each institution in the collaboration. For details on the recruitment sites, the research approvals, and protocols, please consult the online "Supplementary Note" of the original publication [115], as well as the study outline for these publicly available data at dbGaP (`https://www.ncbi.nlm.nih.gov/gap`) under study accession ID phs000094.v1.p1.

**Paper II** The work presented in Paper II is based on analytic formulas and Monte Carlo simulations and does not include any research on human subjects or human data. Hence, no specific ethical approvals are required.

# 4 Summary of main results

## 4.1 Paper I

We developed new methodology to assess PoOxE effects in case-parent triads with or without unrelated control families. We illustrated that PoOxE effects can occur even in the absence of separate PoO or GxE effects. Haplin allows for parallel processing of analyses, and the run time of a genome-wide PoOxE scan is therefore satisfactory. The power to detect a PoOxE effect is approximately 80% using a nominal significance level of 5%, a relative risk ratio of 1.6 ($\text{RR}_{M,2} = 1.6$ and $\text{RR}_{M,1} = \text{RR}_{F,1} = \text{RR}_{F,2} = 1$), a MAF of 0.2, and a total sample size of 2000 case-parent triads with equally sized exposure groups. However, changing the nominal significance level to $5 \cdot 10^{-8}$, a total of 10,000 case-parent triads are needed to reach the same power. Since the PoOxE analysis tests for a second-order interaction effect, a larger sample size is required for the PoOxE test to achieve the same power compared with tests for similar PoO or GxE effects. We also showed that the PoOxE test is asymptotically unbiased. However, when the number of case families is too small in one or several exposure groups, the attained significance level may not match the nominal.

The layout of Table 1 and Table 2 in the published article makes them somewhat difficult to read. The tables, in their submitted versions, are therefore attached in Appendix II.

## 4.2 Paper II

Based on log-linear modeling, we implemented a complete setup for power calculations in Haplin. Statistical power calculations can be performed for child, PoO, maternal, and GxE effects, and an inherent strength of the Haplin framework is the ability to compute power for both single SNPs and haplotypes, either autosomal or X-linked. Moreover, Haplin accommodates family-structure in data, and a wide range of study designs are therefore applicable for power analyses.

In Paper II, we showcased the functionalities for power analysis in Haplin by extensive examples. For the mfc, mc, mfc-mfc, and mc-mc designs, we illustrated that the statistical power to detect a given child effect is identical to that of a maternal effect when adjusting for the possible confounding of the effects with one another. Furthermore, we showed that unrelated control-parent triads do not add extra power to the mfc design when investigating PoO effects.

Statistical power analyses in Haplin can be carried out in two ways, either analytically by using the asymptotic variance-covariance structure of the parameter estimator, or else by using a straightforward Monte Carlo simulation approach. We showed that the two procedures for power calculation provide similar results. For external validation, we further compared the Haplin power calculation module to the power of analysis attained in simulations with EMIM. The consistent results observed between Haplin and EMIM across different study designs and genetic parameterization models confirm the computational accuracy of the statistical inference methods used in both software. They also show that the power calculations in Haplin are applicable to genetic association studies analyzed by either log-linear or multinomial modeling approaches. In summary, the results indicate that Haplin provides a versatile and robust framework for power calculations in genetic association analyses for a broad range of different genetic effects and etiologic scenarios, based on a variety of family-based study designs.

## 4.3 Paper III

In Paper III, we argued for augmenting power analysis with relative efficiency when designing a genetic association study. We introduced a comprehensive framework for relative efficiency estimation and provided insights into how relevant designs compare according to relative efficiency. The methodology was illustrated with analyses of child (regular autosomal), PoO, and maternal effects, using the Pitman efficiency.

Our findings relate to power and efficiency considerations only. For child effects, the c-c design is recommended, and an equal number of cases and controls maximizes the efficiency. For a PoO analysis, optimal efficiency is achieved for the mfc or mc/fc design, depending on the MAF. We also observed that unrelated controls or control families would not increase the power attained by the mfc design, as previously indicated in Paper II. For maternal effects, the results suggest that the mfc design would be an overall good choice when adjusting for child effects, whereas the mfc-c or mc-mc design would be appropriate when adjusting for PoO effects.

In a search for child or PoO effects, we found that an adjustment for maternal genes could cause a substantial loss of power. Hence, we do not recommend including maternal effects in a full GWAS scan for child or PoO effects. As an alternative, we propose additional post-scan analyses to control for the possible confounding.

We also showed that the relative efficiency depends on both the genetic effect in question and the MAF of a given SNP. The results presented are thus subject to the investigated parameter values and should not be interpreted as general guidelines. Furthermore, practical issues should always be taken into consideration, such as the availability of parental genotypes or an appropriate control sample, as well as costs related to recruitment and phenotyping. Nevertheless, relative efficiency is a useful measure for optimizing the study design, and a careful review of relevant designs should be performed as a routine *prior* to performing a GWAS.

# 5 Discussion

The log-linear model in Haplin forms the basis of this thesis, and the new methodologies and developments have been integrated into this framework. However, several other statistical software have also been designed for genetic association analysis, permitting the estimation and testing of genotype relative risk parameters similar to those investigated in Haplin. Although a complete listing and comparison are beyond the scope of this thesis, I will nonetheless briefly review some of the most commonly used tools. I will also mention some of the most prominent power calculation software for genetic association testing. Further, I will discuss some general methodological issues that are relevant to the papers herein. Genetic effects and study designs in genetic association testing are the primary objectives of the thesis. I will, therefore, summarize this section with a few additional remarks on these topics that have not been fully discussed in Papers I—III. Specific discussion points from the individual papers will not necessarily be repeated. Strengths and limitations will be discussed consecutively.

## 5.1 Statistical software for genetic association analyses

The most widely used software for whole-genome association and population-based linkage analyses is PLINK. The original paper about this software [113] has been cited more than 13,000 times according to the Web of Science Core Collection (`apps.webofknowledge.com`; accessed October 9, 2019). PLINK was developed to perform a number of basic, large-scale analyses. It is a computationally efficient tool for data management, quality control, and GWAS. However, few family-based association tests are incorporated in the software, and a complete setup for likelihood-based estimation is lacking. The ability of PLINK to detect PoO effects is limited to an intuitive TDT-like approach, which, as noted in Section 1.8, is not generally statistically valid [50, 98]. Both Haplin and EMIM, which are investigated in this thesis, are able to handle input data from PLINK. The software can therefore be used in combination. For example, PLINK can be used for the initial data management and quality control, whereas Haplin or EMIM can be applied for the actual testing of the genotype relative risk parameters. Note, however, that PLINK holds several functionalities for association testing that are not implemented in Haplin or EMIM, including quantitative trait and sibship analysis.

Another well-established and flexible tool for genetic association analysis of binary and quantitative traits is UNPHASED [64]. A variety of family-based study designs, including sibship data, can be analyzed using UNPHASED, and missing genotype data are handled through direct maximization of the incomplete-data likelihood. UNPHASED can perform haplotype analysis on both autosomal and X-linked data, and modification of haplotype relative risks according to the parent of origin has also been implemented. In addition, UNPHASED can run a sliding-window analysis over a selection of markers, similar to Haplin. Although a sliding-window approach is convenient for GWAS analysis, UNPHASED is not designed for this purpose. Dudbridge, the author of UNPHASED, suggests that GWAS packages (e.g., PLINK) should be used for the initial quality control and analysis [118]. UNPHASED could then be applied to verify promising findings in situations where its approaches are more efficient, such as family-based studies with incomplete parental information or haplotype analysis. A comparison of family-based methodologies for assessing haplotype effects on the X chromosome showed that Haplin has more consistent Type I error rates and better power than UNPHASED, even when HWE is not fulfilled [119]. However, the comparisons were conducted under simulated scenarios corresponding to X-chromosome models that are available in Haplin, thus optimizing the performance of Haplin.

Another genetic analysis software, Mendel, performs likelihood-based statistical analysis of binary and quantitative traits for a wide range of genetic problems, covering both parametric linkage in large pedigrees and genome-wide association analysis of rare alleles [120, 121]. Mendel encompasses several options for association testing, including the maternal-fetal genotype (MFG) incompatibility test for assessing both child and maternal effects as well as mother-child interaction effects [57]. The MFG test for binary traits is also implemented in EMIM, and a comparison of EMIM and MENDEL showed similar inference. However, EMIM is faster and provides an easier implementation of various parameterization models [105]. Although the MFG test was developed using a log-linear modeling approach for case-parent triads, it has not yet been implemented in Haplin. Thus, power calculations and relative efficiency measures for the MFG test are currently lacking, and the incorporation of the MFG test would have been advantageous to this thesis.

Similar to Haplin, LEM (Log-linear and event history analysis with missing data using the EM algorithm) [122–124] is also based on log-linear modeling and designed to analyze child, PoO, and maternal effects in family triads. Howey and Cordell [105] demonstrated that the inference provided by EMIM and LEM is similar, which is as expected due to the mathematical equivalence between the multinomial and log-linear model [104, 114]. However, EMIM is considerably faster. For PoO and maternal effects, Connolly and Heron [50] reviewed different statistical methods for binary traits and compared them according to Type I error rates, statistical power, and suitability for studying different etiologic scenarios. The multinomial model in EMIM was recommended because EMIM has the most consistent Type I error rate, attains the strongest power, is easy to implement, and offers additional flexibility. Regrettably, Haplin was not included in that review, but the comparisons between Haplin and EMIM in Paper II showed similar inference across multiple genetic effects and study designs. We have also conducted extended analyses using a lower significance threshold ($\alpha = 10^{-4}$) than what was demonstrated in Paper II, and the close correspondence between EMIM and Haplin still holds (results not shown). Hence, EMIM and Haplin are both reliable and versatile approaches for performing genetic association analysis based on genotype data from a wide range of child-parent configurations, offering, to a broad extent, similar functionalities. The advantage of Haplin is that it is able to examine both X-linked and GxE effects. EMIM, on the other hand, allows a variety of likelihood assumptions other than HWE, such as mating symmetry [99], parental allelic exchangeability [83] and a "conditional on parental genotype" model [102]. The different likelihood assumptions will be discussed further in Section 5.2.2.

There are also several other software that deserve to be mentioned. TRANSMIT [125] was one of the first software to test for association between a genetic marker and a disease trait by investigating the transmission of markers from parents to affected offspring. It can handle missing parental genotypes as well as the transmission of multi-locus haplotypes, even when the haplotype phase is unknown. The TRIad Multi-Marker (TRIMM) method [126] was designed to detect risk-related haplotypes by applying multiple SNPs from case-parent triads directly without having to infer haplotypes. TRIMM offers a non-parametric approach for testing multiple SNPs simultaneously. It can accommodate deviations from HWE, population structure, and non-negligible rates of recombination, and the methodology can be used to investigate child, PoO, and maternal effects. Although TRIMM is better at detecting associations dominated by a single SNP or haplotype, Haplin performs better when several risk-associated haplotypes are involved [127]. The Bioconductor package `trio` [128] in **R** specializes in genome-wide analyses of case-parent triad data, and both TAT [99] and the parent-of-origin likelihood ratio test (PO-LRT) [98] have been included to detect PoO effects. I will end this section by acknowledging GenABEL [129], an **R** package designed to handle GWAS data in a memory-efficient way, facilitating both data management and quality control in **R** as well as GWAS analysis. Up until 2018, Haplin depended on GenABEL to convert .ped files to Haplin format. Unfortunately, GenABEL was discontinued and removed from the CRAN repository in May 2018 due to lack of maintenance.

### 5.1.1 Power calculation software

Genetic Power Calculator (GPC) is an easy-to-use tool to calculate statistical power for linkage and association mapping [92]. The paper by Purcell et al. [92] has been cited nearly 1650 times according to the Web of Science Core Collection (`apps.webofknowledge.com`; accessed October 9, 2019), reflecting its valuable contribution to the research community. Another well-known power calculation software for genetic studies is Quanto, which offers power and sample size computations for child effects, GxE, and gene-gene interactions [130–132]. GPC and Quanto are both based on closed-form analytic power formulas. They can perform power calculations for both quantitative and binary traits, and power analysis of sibship data is also incorporated. Nevertheless, only a limited number of study designs are available that accommodate parental information and family structure. Moreover, power analyses of PoO and maternal effects have not been implemented, and power calculations involving X-linked markers or haplotypes are not available in the modules implemented to date in either software. Unfortunately, neither GPC nor Quanto seem to have been updated in recent years. Notably, several of the modules in GPC are undocumented and unsupported, and the latest version of Quanto was released in 2009.

The PBAT software [133, 134] includes a unified approach to the family-based association test (FBAT) [135], which is a generalization of the TDT. It incorporates nearly all of the features of the preceding FBAT package [136] but also provides power calculation functions for binary and quantitative disease traits, thus accommodating a wide range of family- and population-based study designs with the ability to handle missing parental genotypes [137–139]. Similar to Haplin, PBAT includes functionalities for verifying the analytical power calculations by Monte Carlo simulations. Although a thorough comparison between the power functionalities of Haplin and PBAT would be useful, PBAT and FBAT have been incorporated in the commercialized Golden Helix Foundation software (`https://www.goldenhelix.com`), thus preventing the widespread use of their tools in academic research.

The analytical power calculations in Haplin are general; we apply the asymptotic normal distribution of the log-transformed relative risk and relative risk ratio parameters and use the non-centrality parameter for the test statistic. The power functions in Haplin are implemented as part of a unified analysis setup. This makes it easy to extract the variance-covariance matrix needed to compute the non-centrality parameter in different scenarios. As additional methods of analysis are developed and implemented in Haplin, corresponding power calculations can readily be incorporated, both analytically and through simulations. Hence, further advancements of power functionalities are more easily achievable in Haplin than in independent power calculation software that are based on closed-form analytic equations for each situation (i.e., not integrated as part of a general analysis framework).

Our power calculations are, nevertheless, restricted to binary disease traits. Even though power calculations for quantitative traits can also be performed based on the non-centrality parameter for the Wald test statistic, the quantitative-trait non-centrality parameter cannot be calculated within the current Haplin setup. In Haplin, quantitative traits such as birth weight or gestational length can be analyzed by dichotomizing the outcome variable. For example, gestational length may be assessed by assigning the cut-off at preterm birth, defined as birth prior to 37 completed weeks of gestation [140]. However, the dichotomization of quantitative variables may cause several problems, and valuable information can be lost [141]. For analytical power calculations of quantitative traits, please consult the recent publication of Wang and Xu [111].

## 5.2 Methodological considerations and limitations

For our PoOxE calculations in Paper I, we assumed independence between exposure and the child's genotype conditional on parental mating type. However, I will show here that this constraint can be relaxed, as previously derived in our recent paper [61]. The reliance on the HWE assumption has been a matter of some debate, and several alternative likelihood assumptions have been introduced in the literature. A short summary of the most important parameterizations will be provided below.

Owing to the excessive number of SNPs being tested in a GWAS, issues of multiple comparisons are of major concern in whole-genome association testing. Although methods for handling multiple testing are beyond the scope of this thesis, a few general remarks will be provided in Section 5.2.3.

### 5.2.1 The assumption of conditional independence between exposure and child genotype given parental genotypes

The standard log-linear model describes the probability of the case-parent triad genotype, conditional on the child being a case, and can be parameterized as

$$P(M, F, C|D) = \frac{P(D|M, F, C)P(M, F, C)}{P(D)}.$$

If we assume that information about the parental genotypes is irrelevant for the disease penetrance when the genotype of the child is known, we have that

$$P(M, F, C|D) = \frac{P(D|C)P(M, F, C)}{P(D)},$$

as explained in Section 3.1. With a categorical exposure variable, $E$, included, the analogous parameterization is

$$P(M, F, C, E|D) = \frac{P(D|C, E)P(E|M, F, C)P(M, F, C)}{P(D)},$$

assuming that the parental genotypes for the disease penetrance are irrelevant when the child's genotype and exposure status are known. To estimate GxE effects in case-parent triads, a standard constraint is independence between $C$ and $E$ conditional on parental genotype, i.e., $P(E|M, F, C) = P(E|M, F)$ [79, 142]. Note that $(M, F, C)$ here denotes the unordered triad type, as opposed to the strict ordering used in Section 3.1. The sampling model can be expressed as

$$\begin{aligned} P(M, F, C, E|D) &= \frac{P(D|C, E)P(E|M, F)P(M, F, C)}{P(D)} \\ &= \frac{P(D|C, E)P(M, F|E)P(C|M, F)P(E)}{P(D)}, \end{aligned}$$

where $P(D|C, E)P(M, F|E)P(C|M, F)$ corresponds to a stratum-specific log-linear model. Since $P(E)$ and $P(D)$ are constant within a stratum, the log-linear model can be fitted directly within each stratum (see Section 3.1).

When estimating PoOxE effects (Paper I), the sampling model can be parameterized as

$$P(M,F,C_{jl},E|D)=\frac{P(D|C_{jl},E)P(E|M,F,C_{jl})P(M,F,C_{jl})}{P(D)},$$

where $C_{jl}=A_jA_l$ denotes that allele $A_j$ is inherited from the mother and allele $A_l$ is inherited from the father. Assuming that $P(E|M,F,C)=P(E|M,F)$, we could proceed as for the child effects. However, to estimate the ratio $\mathrm{RRR}_j=\mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$ within each stratum, a less stringent assumption would suffice [61]. With the constraint that

$$P(E|M,F,C_{jl})=P(E|M,F,C_{lj})=P(E|M,F,C), \tag{7}$$

i.e., the alleles of the child may affect the exposure directly, even within parental mating types, but the effect should not depend on parental origin, we have that

$$P(M,F,C_{jl},E|D)=\frac{P(D|C_{jl},E)P(M,F|E)P(C|M,F,E)P(E)}{P(D)}\cdot\frac{P(C_{jl}|M,F)}{P(C|M,F)},$$

where the latter fraction depends on Mendelian inheritance. For the standard evaluation in Haplin, the log-linear model is fitted within each exposure stratum. However, since $P(C|M,F,E)$ may depend on both $E$ and the (unordered) $C=A_jA_l$, the separate within-stratum estimates of $\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ may be biased. Nevertheless, for the ratio $\mathrm{RRR}_j=\mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$ obtained in each stratum, the bias cancels out if Eqn. (7) holds true. The different assumptions required for GxE and PoOxE analyses for a variety of study designs have previously been described in our recent paper [61].

An example where the conditional independence assumption $P(E|M,F,C)=P(E|M,F)$ could fail, possibly biasing the GxE estimate, is if the variant allele itself directly affects an individual's propensity for the exposure, either through appetite or aversion. For instance, an individual's reluctance toward excessive alcohol intake may be associated with a genetic variant that slows the detoxification of alcohol [143]. However, unless this mechanism depends on the parent of origin, the PoOxE estimate would still be valid.

### 5.2.2 Deviations from HWE

The triad population frequencies $P(M, F, C) = P(A_iA_j, A_kA_l)$ can be parameterized in different ways. The basic implementation in Haplin employs haplotype-frequency parameters under the HWE assumption. Thus, $P(M, F, C) = p_ip_jp_kp_l$, where $p_i$ is the population frequency for allele $A_i$, with the constraint that $\sum_i p_i = 1$. While HWE is a reasonable assumption in random mating populations, it is disputed in populations with substructures. The case-parent triad design inherently protects against population stratification, but some of this protection is lost if the HWE assumption is not satisfied. To avoid the HWE constraint, Wilcox et al. [52] and Weinberg et al. [99] made the less strict assumption of "mating symmetry" and introduced six mating type parameters, $\mu_1 - \mu_6$, as shown in Table 7. However, the number of parameters can be reduced in the presence of inbreeding or population stratification, situations which typically cause an excess in the observed proportion of homozygotes from what would be expected under HWE [144]. Gjessing and Lie [16] suggested modeling such deviations by the triad frequencies $P(M, F, C) = p_ip_jp^*_{ij}p_kp_lp^*_{kl}$, where $p^*_{ii} = p^*_i$ for each homozygote and $p^*_{ij} = 1$ for all heterozygotes. This parameterization has been implemented as an addition to the standard HWE model in Haplin, but it is not yet available in the official version.

Several other constraints have been proposed in the literature. For example, the "conditional on parental genotypes" (CPG) introduces as many as nine mating type stratification parameters, $\mu_1 - \mu_9$ (Table 7) [83, 101, 102]. This model should be more robust to departures from mating symmetry or HWE but loses statistical power compared with corresponding models with fewer parameters [105]. The "parental allelic exchangeability" (PAE) assumption asserts that the four alleles carried by a pair of parents in the source population are randomly distributed among them [83, 126]. In the context of the parameterization of Wilcox et al. [52] and Weinberg et al. [99], this corresponds to setting $\mu_4 = \mu_3$ (Table 7). This restriction is slightly stronger than mating symmetry but considerably weaker than HWE since it still allows populations with substructures.

The HWE assumption reduces model complexity. Hence, HWE simplifies computations and improves the computational efficiency in Haplin. It also facilitates haplotype reconstruction; the parameterization of Table 4 in Section 3.1 is readily extended to haplotype analysis. Such extensions would become more cumbersome with the parameterization outlined in Table 7, where the multinomial is categorized according to the number of copies of the variant allele carried by the mother, father, and child. While fewer parameters lead to a more constrained model, the statistical power is increased, provided the model is correct. As a consequence, a power analysis in Haplin would typically overestimate the power of log-linear or multinomial models adopting the mating type parameter approach. Comparisons of different constraints in EMIM show that the power to detect an association decreases as one makes less restrictive but potentially more robust constraints [105, Figure 2]. The decrease in power is more pronounced for child and maternal effects than for the EMIM maternal imprinting effects. The loss of power is also greater at lower significance thresholds. Nonetheless, most researchers would not have sufficient knowledge at the planning stage of a study to be able to realistically specify possible configurations of mating type parameters.

Since the default implementation in Haplin uses the HWE assumption, an analysis scheme should always include a strategy for investigating large deviations from HWE. Haplin performs a chi-squared test for HWE on all SNPs as an automated part of all analyses. Thus, as a routine, top hits from a GWAS analysis in Haplin should be checked post hoc to prevent spurious associations caused by, for instance, population stratification, genotyping errors, or deviations from Mendelian transmission.

**Table 7:** Population frequencies of case-parent triads for a diallelic SNP. A comparison of different constraints

| Row number | Triad genotype (M,F,C)* | CPG | Mating symmetry | PAE | HWE |
|---|---|---|---|---|---|
| 1 | 2,2,2 | $\mu_1$ | $\mu_1$ | $\mu_1$ | $p_2^4$ |
| 2 | 2,1,2 | $\mu_2$ | $\mu_2$ | $\mu_2$ | $p_1p_2^3$ |
| 3 | 2,1,1 | $\mu_2$ | $\mu_2$ | $\mu_2$ | $p_1p_2^3$ |
| 4 | 1,2,2 | $\mu_3$ | $\mu_2$ | $\mu_2$ | $p_1p_2^3$ |
| 5 | 1,2,1 | $\mu_3$ | $\mu_2$ | $\mu_2$ | $p_1p_2^3$ |
| 6 | 2,0,1 | $\mu_4$ | $\mu_3$ | $\mu_3$ | $p_1^2p_2^2$ |
| 7 | 0,2,1 | $\mu_5$ | $\mu_3$ | $\mu_3$ | $p_1^2p_2^2$ |
| 8 | 1,1,2 | $\mu_6$ | $\mu_4$ | $\mu_3$ | $p_1^2p_2^2$ |
| 9 | 1,1,1 | $2\mu_6$ | $2\mu_4$ | $2\mu_3$ | $2p_1^2p_2^2$ |
| 10 | 1,1,0 | $\mu_6$ | $\mu_4$ | $\mu_3$ | $p_1^2p_2^2$ |
| 11 | 1,0,1 | $\mu_7$ | $\mu_5$ | $\mu_4$ | $p_1^3p_2$ |
| 12 | 1,0,0 | $\mu_7$ | $\mu_5$ | $\mu_4$ | $p_1^3p_2$ |
| 13 | 0,1,1 | $\mu_8$ | $\mu_5$ | $\mu_4$ | $p_1^3p_2$ |
| 14 | 0,1,0 | $\mu_8$ | $\mu_5$ | $\mu_4$ | $p_1^3p_2$ |
| 15 | 0,0,0 | $\mu_9$ | $\mu_6$ | $\mu_5$ | $p_1^4$ |

* Number of copies of the $A_2$ allele, e.g., 2,2,2 denotes the triad genotype $A_2A_2, A_2A_2, A_2A_2$

CPG: conditional on parental genotypes; PAE: parental allelic exchangeability;

HWE: Hardy-Weinberg equilibrium

$\mu_1 - \mu_9$ are mating type stratification parameters;

$p_1$ and $p_2$ are the allele frequencies corresponding to $A_1$ and $A_2$, respectively, i.e., $p_1 + p_2 = 1$.

The sum of the relative-frequency parameters across the 15 categories is constrained to 1

### 5.2.3 Multiple testing issues

In Paper II, the analyses were carried out using a nominal significance level of $\alpha = 0.05$, assuming that there is only a single locus (i.e., one hypothesis) under investigation. As there are several hundred thousand SNPs in a GWAS, it is crucial to correct for multiple comparisons. The most common way of adjusting for multiple testing is by using the Bonferroni corrected significance level, defined by $\alpha_{\text{Bonferroni}} = \alpha/m$, where $\alpha$ is the family-wise significance level and $m$ is the number of (independent) hypotheses being tested. The Bonferroni method controls the family-wise error rate (FWER) at level $\alpha$, i.e., guarantees that the probability of making at least one Type I error does not exceed the given significance threshold. Another well-established method for controlling the FWER is the Šidák correction, defined by $\alpha_{\text{Šidák}} = 1 - (1-\alpha)^{1/m}$, which is marginally less conservative than the Bonferroni adjustment. Both of these methods are easily handled by the Haplin power calculation modules, by modifying the significance level to the appropriate threshold. Power analyses investigating different FWER thresholds (more relevant to a GWAS) were performed in Paper I.

The commonly used Bonferroni and Šidák corrections are overly conservative when the tests are not statistically independent. This would be the situation in a GWAS, since a large proportion of SNPs are in LD. Modifications have been suggested to allow correlations between adjacent SNPs, for example by evaluating the effective number of independent tests [8]. The widely adopted genome-wide significance threshold of $5 \cdot 10^{-8}$ is equivalent to a Bonferroni adjustment of $10^6$ tests, assuming that dependencies between neighboring SNPs are so strong that a full GWAS search corresponds to conducting $10^6$ tests, regardless of the number of SNPs actually being tested [93].

Instead of controlling the FWER, however, another approach entails controlling the false discovery rate (FDR), i.e., the expected proportion of true null hypotheses among the null hypotheses that have been rejected [145]. If the FDR is controlled at the 0.05 level, this approach ensures that no more than 5% of the reported significant findings will be false positives. The FDR method is less conservative than the FWER corrections, resulting in better power to reject the null hypothesis when the alternative hypothesis is true. Thus, sample size estimation for a specified number of true rejections, while controlling the FDR at a given threshold, would be of great importance in genetic studies. However, these calculations would also depend on the effect sizes among the true positives [146], which is normally unknown.

Although sample size calculations under FDR control could be obtained through simulations, the incorporation of such procedures is beyond the scope of this thesis. They would be a useful extension to the power calculation module in Haplin in future developments of the software.

## 5.3 Genetic effects and study designs

The GWAS is used to identify associations between genetic markers and traits in samples from populations, with the primary aim of increasing knowledge of disease biology [3]. A straightforward GWAS search for child effects may discover several markers associated with the disease. However, even when a marker allele is the actual disease-causing allele, the translation from a GWAS finding to biology is fraught with difficulties, as one may not understand the underlying mechanisms causing the statistical association. For example, if $\mathrm{RR}_M = 2$ and $\mathrm{RR}_F = 1$, an attenuated association might still be detected in a search for child effects, although the effect is maternally inherited. Hence, statistical methodologies that can distinguish between various casual models are essential not only for the identification of new disease loci but also for advancing the understanding of the biological mechanisms involved [26].

The power to detect complex scenarios, e.g., interaction effects, in a full GWAS analysis is generally limited, and a strategy for selecting candidate genes would help to reduce the number of tests. For PoO effects, the search could be limited to imprinted genes, or, alternatively, to top hits from a GWAS scan for child effects. However, these types of candidate-gene approaches have suffered from poor replication rates in follow-up studies [30], and the optimal strategy is not known a priori.

Numerous studies have aimed to detect GxE and PoO effects separately for a large number of traits and diseases [40, 116, 117]. However, because maternal environmental factors affecting methylation patterns might also influence the effects of maternally and paternally inherited alleles in unequal measure, it is reasonable to assume that the joint interaction effect may also affect the risk of a complex disease. In 2011, Wang et al. [147] developed a logistic regression approach for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. Although the need to develop methods for analyzing PoOxE effects has been warranted [26, p. 616], the procedure has had relatively little impact in terms of citations (cited six times in the Web of Science Core Collection (`apps.webofknowledge.com`; accessed November 10, 2019)). Three notable limitations of their approach are the restriction of study design, the inability to account for maternal effects, and the lack of implementation in software. Our recently developed PoOxE test (Paper I) is based on log-linear modeling and can readily be adapted to accommodate these limitations. The ability to integrate a wide range of family-based study designs improves estimation by using all the available information. Analyses are not restricted to a fixed design; genotype data from various child-parent configurations can be combined, such as a mixture of case-parent triads and case-mother dyads, possibly supplemented by unrelated cases and controls. Furthermore, the inclusion of the methodology in Haplin is a prerequisite for ease of use, and researchers can readily apply our functions to investigate PoOxE effects in their own data.

When investigating a GxE or a PoOxE effect, the environmental exposure may refer to an individual's exposure to, for example, alcohol, smoking, diet, or exercise. However, when studying birth defects, the environmental exposure will typically refer to a maternal exposure. It may also refer to a stratification factor, such as ethnicity or study site. In Paper I, the exposure of interest was maternal smoking during the periconceptional period. However, in a few instances, the two strata referring to paternally inherited alleles were labeled as "exposed" and "unexposed" fathers. Although the strata were correctly categorized, this labeling was imprecise as the exposure status should be referring to that of the mother of their child.

In a recent paper [61], we used Haplin to search for statistical interactions between a SNP allele and DNA methylation (GxMe) and between a PoO effect and DNA methylation (PoOxMe), treating the methylation level as the exposure of interest. This can be viewed as a direct application of the GxE and PoOxE approaches in Paper I, and the same constraints would therefore apply. A GxMe search relies on the conditional independence assumption between exposure (methylation) and the child's genotype given the genotypes of the parents. This assumption may not hold when the methylation levels at a CpG site are directly affected by a nearby SNP, which is the situation for methylation quantitative trait loci (meQTLs) [148–150]. A PoOxMe investigation relies on a less stringent assumption (Eqn. 7) and is therefore more likely to be valid. A PoOxMe investigation might also be biologically intriguing. For instance, since a PoO effect may result from imprinting, and since imprinting may occur through differing methylation levels depending on parental origin, we might anticipate that methylation levels at nearby CpGs could actively affect the magnitude of the PoO effect [61, 151].

The definition of PoO effects in the literature is somewhat ambiguous in that both genomic imprinting and trans-generational (e.g., maternal) effects have been described as parent-of-origin effect types [50, 51]. The parameterization of PoO effects is relatively complex, and various models have been proposed that allow for different interpretations, as reviewed by Ainsworth et al. [104]. The PoO effect investigated in Papers I—III pertains to the parameterization in Table 5 and is defined by the ratio $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F$. However, an assessment of maternally ($\mathrm{RR}_M$) or paternally ($\mathrm{RR}_F$) inherited PoO effects might also be of interest. In Haplin, both $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are estimated freely, and individual tests for the null hypotheses $\mathrm{RR}_M = 1$ and $\mathrm{RR}_F = 1$ are performed. Figure 7 (see page 82) shows the relative efficiency for testing the hypotheses $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F = 1$ (PoO effect, blue line), $\mathrm{RR}_M = 1$ (the effect of alleles of maternal origin, orange line), and $\mathrm{RR}_F = 1$ (the effect of alleles of paternal origin, green line), where **i)** compares the case-mother dyad design relative to the case-parent triad design, **ii)** compares the case-father dyad design relative to the case-parent triad design, and **iii)** compares the case-father dyad design relative to the case-mother dyad design. The allele frequency corresponds to the risk increasing allele.

For PoO effects under $H_0$ (Figure 7a), we observe that the efficiency of the case-mother and case-father dyad designs exceeds that of the case-parent triad design for allele frequencies less than 0.25 or above 0.75. These findings are in agreement with the observations in Paper III. However, somewhat counter-intuitive, we see that case-mother dyads provide better efficiency than case-parent triads and case-father dyads when testing for the effect of paternally derived alleles. By symmetry, we also observe that case-father dyads provide better efficiency than case-parent triads and case-mother dyads when testing for the effect of maternally derived alleles. Moreover, under $H_0$, the case-mother and case-father dyad designs appear to be equally efficient for testing the ratio $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F = 1$. However, when $\mathrm{RR}_M > 1$ and $\mathrm{RR}_F = 1$ (Figure 7b), we observe that the case-mother dyad design attains better efficiency for allele frequencies less than 0.5, whereas the case-father dyad design attains better efficiency for allele frequencies above 0.5. A similar discussion (possibly with slightly different parameterizations for the effect of maternally and paternally derived alleles) was also made by Howey et al. [106]. Most of their findings are in agreement with those of Figure 7, but a few observations might seem to go in the opposite direction. Our results have been thoroughly checked through simulations in both Haplin and EMIM, which are consistent, and the inconsistencies might therefore be due to different methods, tests, or parameterization models used in the simulations or analyses. Nonetheless, the source of the discrepancies has not yet been identified. Howey et al. [106] also conclude that the case-parent triad design provides better power than the case dyad designs. Although this is true when comparing an equal number of case families (e.g., comparing 500 case-mother or case-father dyads with 500 case-parent triads), their conclusion does not take relative efficiency into account.

While no rare disease assumption is necessary for relative risk estimation in case-parent triads [52], this assumption is required to incorporate unrelated and truly unaffected controls in the Haplin framework. The log-linear model in Haplin assumes that the control sample comprises a random sample from the population, i.e., that the controls are of unknown disease status. Given that the control sample is truly unaffected, this corresponds to making a rare disease assumption (see Section 3.1). Thus, if the disease is rare, either unaffected or population-based controls can be used [104]. By contrast, for a common disease, the log-linear structure would be lost if the controls are truly unaffected, but one may proceed with the relative risk estimation if the controls are randomly sampled from the population [140]. This distinction is essential because the controls are used to aid the estimation of population allele frequencies. If the disease is common, these frequencies might be biased if they are fitted conditional on (unaffected) disease status in the offspring.

In Paper II, we found that unrelated control families would not improve the power obtained by the case-parent triad design alone when estimating PoO effects. This was also demonstrated in Paper III, where the relative efficiency of the hybrid designs decreases when the ratio of control families to case families increases or when the number of genotyped individuals within a control family increases. Nevertheless, independent control-parent triads may still be useful because they allow estimation of the main effects of an exposure. Moreover, unrelated control families are necessary to check key assumptions underlying the case-parent triad design (e.g., HWE, mating symmetry, and Mendelian transmission), and to account for false positive findings. For instance, Eqn. (7) would not be needed for the assessment of PoOxE effects if $P(E|M, F, C)$ could be estimated directly among control-parent triads. Alternatively, unbiased within-stratum estimates of $\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ could be obtained by including control-parent triads in a hybrid analysis [61].

Family-based study designs facilitate the estimation of genetic effects without bias from population stratification, and various child-parent configurations have been interrogated and compared herein. However, the case-sibling design, in which each case is matched to one or more unaffected siblings, has not yet been discussed. For case-sibling studies, within-family dependencies can be accounted for by applying conditional logistic regression [152]. Alternatively, the case-sibling data can be considered as nuclear family data with parents missing by design. Under a rare disease assumption, this information can be incorporated into a log-linear framework, in which the missing-parents likelihood can be maximized via the EM algorithm [153, 154]. Case-sibling studies are of particular value for diseases with late onset when parents may not be available, although parental information is still needed to estimate PoO or maternal effects. Moreover, with information on exposure status of unaffected siblings, estimation and testing of environmental influences are feasible, which cannot be done with the case-parent triad design alone. However, estimation of main exposure effects would not be possible in twin studies if the environmental factor refers to a maternal exposure during pregnancy. When investigating child effects, the case-sibling study design has less power than both the case-parent triad and the case-control designs [155]. Assuming a multiplicative dose-response relationship, power calculations in Quanto [130–132] show that the use of unmatched case-control pairs is approximately twice as efficient as case-sibling pairs and that the case-parent triad design is 4/3 times as efficient as the case-sibling design. These relative efficiency estimates are independent of the MAF. For a rare phenotype, the case-sibling design with infinitely many siblings would provide the same information as the case-parent triad design [154], and the inclusion of unaffected siblings would thus only improve estimation if one or both parents are missing.

*Literature review completed December 2019.*

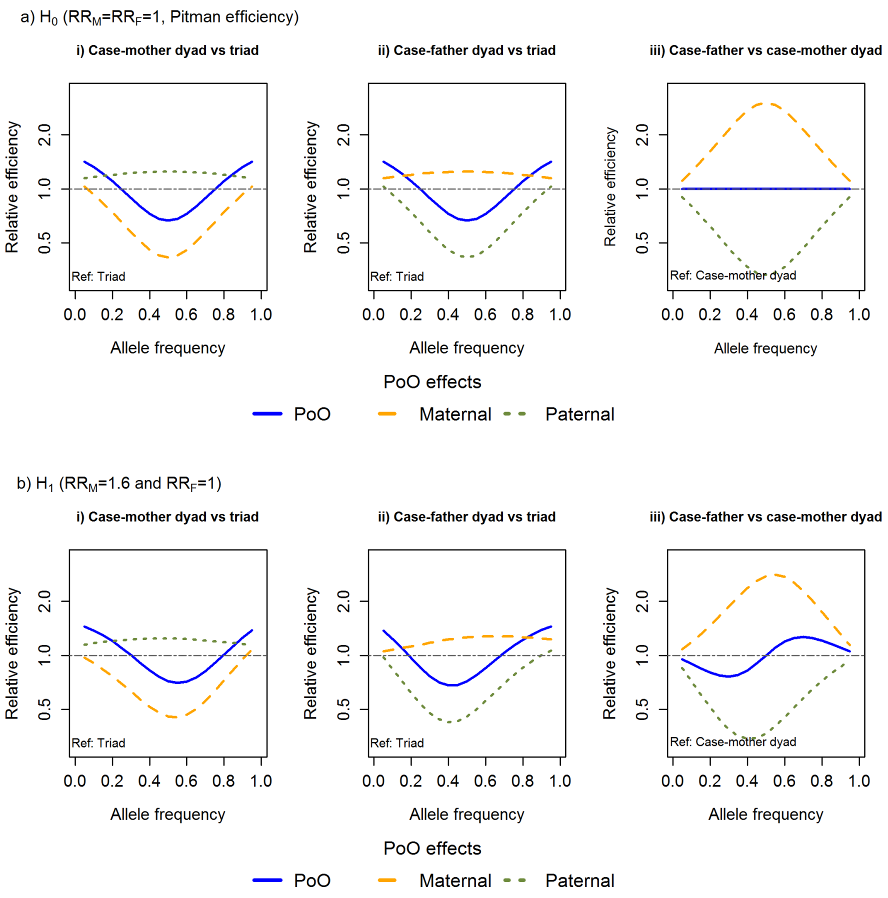

**Figure 7:** Relative efficiency for testing the hypotheses RRR $= \mathrm{RR}_M/\mathrm{RR}_F = 1$ (blue line), $\mathrm{RR}_M = 1$ (orange line) and $\mathrm{RR}_F = 1$ (green line), comparing the case-mother dyad, case-father dyad, and case-parent triad designs. The allele frequency corresponds to the risk increasing allele

# 6 Concluding remarks and further perspectives on GWAS discoveries

Genetic epidemiology aims to study the contribution of genetic risk factors, as well as their interactions with environmental exposures, in determining disease etiology in families and populations. A lot of time and effort have been invested in examining genetic susceptibility to disease, but despite decades of extensive research, the genetic basis of complex diseases remains largely unknown. This underscores the need to interrogate etiologic disease mechanisms other than child effects alone, and we have here developed new methodology for assessing PoOxE effects (Paper I). Although PoOxE effects are likely to explain only a small proportion of the unknown genetic architecture of a trait, a PoOxE search may also be useful for distinguishing between different patterns of gene expression, thus aiding biological interpretation. Since an insignificant test can stem from both the absence of an effect and a lack of statistical power, a sizable fraction of the unknown genetic etiology might also be explained by poorly designed and underpowered studies that are unable to capture most of the genetic variants underlying a trait. To address these shortcomings, we developed a comprehensive setup for power and sample size calculations (Paper II) and used these as building blocks for comparing relevant study designs in terms of relative efficiency (Paper III).

Owing to a general lack of software implementation, it has been difficult to plan, analyze, and interpret genetic studies. The implementation of methodology in Haplin has, therefore, been a priority, and models accommodating family-based data have been a primary concern. The extensive framework for analysis and power calculation in Haplin facilitates not only the analysis of genetic data but also the planning stage of the study, with the aim of making the most out of the available resources. Statistical methodologies that are able to differentiate between various casual models are essential for advancing the field of complex trait research. The establishment of approaches that integrate epigenetic and genetic data is still in its infancy [61, 156–159], and the ability to incorporate methylation and exposure data will provide further opportunities to explore the GWAS design. Nevertheless, challenges remain as to how identified loci can be studied for mechanisms, especially since most identified markers themselves do not cause the disease. Upscaling of fine-mapping technologies and strategies is paramount [160], and replacing SNP arrays with whole-genome sequencing is a natural next step [3].

Large clinical and population-based biobanks and national health registries continue to create new opportunities for genetic, epidemiological, and clinical research worldwide. Sharing of genetic data has facilitated novel research and discoveries. The UK Biobank provides publicly available genetic data on more than 500,000 participants, along with a large collection of phenotypic and health-related information [161]. In Norway, the ongoing Norwegian Mother, Father and Child Cohort Study (MoBa) has genotyped random subjects from large clinical and population-based biobanks and national health registries [162, 163]. To date, it contains information on 11,000 case-parent triads. Larger studies will enable the identification of new loci with smaller effect sizes. It will also allow the detection of variants with lower frequencies [164]. However, for rare-variant associations of a complex disease with low population prevalence, new discoveries will be restricted by the limited number of cases. To test for associations of rare variants, so-called burden or collapsing tests have been introduced, in which rare-variant information in a region is combined into a genetic score or a summary dose variable [165]. With whole-genome sequencing data, such methods can be further improved to increase the statistical power [3, 166, 167].

The architecture underlying complex diseases is multifactorial, consisting of numerous risk loci, structural variants or other forms of genomic variation, intricate gene-gene and gene-environment interactions, as well as epigenetics. Most effect sizes reported are small, and the identification of significant markers is largely dependent on sample size [168]. Polygenic risk scores are routinely used to quantify the cumulative genetic effects among a collection of markers [169]. Each single variant may show a small effect individually, but when analyzed combined, they can be used to identify individuals at higher risk for a given disease [91]. When the sample size is limited, polygenic risk scores can be useful for association testing and for demonstrating a genetic basis even when no single markers alone reach the level of significance in a GWAS [170]. By examining interactions between polygenic risk scores and environmental factors, the power to detect GxE effects can be improved [171]. As the sample size increases, polygenic risk scores can also be used to construct valuable risk prediction models [172–175].

Resolving the multifactorial architecture underlying complex diseases seems like a never-ending task, and it might be just that. Nevertheless, since the introduction of the GWAS design more than two decades ago [27], remarkable discoveries in human genetics have been made, ranging from the identification of genes and loci to a better understanding of the biological pathways involved in complex disease. The substantial improvements of high-throughput technologies and systems approaches have facilitated the translation of GWAS discoveries to biology and treatments. With whole-genome sequencing data, together with detailed phenotypic and -omics data on millions of individuals, new discoveries will continue to improve diagnosis, prognosis, prevention, and treatment. However, the question of whether precision medicine will become the paradigm of health care in the near future remains a matter of debate [3, 168, 176–179].

# 7 Software, electronic database information, and availability

### Haplin

Haplin [16] is implemented as a standard package in the statistical software **R** [107]. It can be installed from the official **R** package archive, CRAN (`https://cran.r-project.org`), from which also the source code for the Haplin functionalities is available. For a thorough description of the Haplin functions and their arguments, please consult our website at `https://people.uib.no/gjessing/genetics/software/haplin`.

**Notes:** Up until May 2018, Haplin depended on GenABEL to store and handle GWAS data. A new and extensive data storage system was developed by Julia Romanowska and introduced in Haplin Version 7.0.0. As a result, minor changes to the Haplin commands presented in the Supporting Information (S1) of Paper I are needed to run the PoO, GxE, and PoOxE analyses. All updates are documented on the **R** help page and on the Haplin webpage.

### PREMIM/EMIM

Information on PREMIM and EMIM [104, 105] is available from `https://www.staff.ncl.ac.uk/richard.howey/emim`.

### CPO data

The GWAS dataset is available from the dbGaP database (`https://www.ncbi.nlm.nih.gov/gap`) under study accession ID phs000094.v1.p1. Details have been provided in the original publication [115].

# 8 Errata

## Paper I

- In Section 3.1, the GxE effect of SNP rs470563 has a $p$-value of $4.5 \times 10^{-4}$, not $4.5^{-4}$.
- In Appendix A.2, $\chi^2_\alpha(r)$ was incorrectly defined as the $\alpha$ quantile of the chi-squared distribution with $r$ degrees of freedom. However, the correct definition is the "upper-$\alpha$ quantile", as defined in Section 3.4.1 in this thesis.

## Paper II

- In Additional file 1, $\boldsymbol{n}$ should be defined as the $q \times 1$ vector $\boldsymbol{n} = [n_1, ..., n_q]^T$.

# References

[1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–753.

[2] Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*. 2010;363:166–176.

[3] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101:5–22.

[4] Sivertsen Å, Wilcox AJ, Skjærven R, Vindenes HA, Åbyholm F, Harville E, et al. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *BMJ*. 2008;336:432–434.

[5] Grosen D, Chevrier C, Skytthe A, Bille C, Mølsted K, Sivertsen Å, et al. A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. *J Med Genet*. 2010;47:162–168.

[6] Grosen D, Bille C, Pedersen JK, Skytthe A, Murray JC, Christensen K. Recurrence risk for offspring of twins discordant for oral cleft: a population-based cohort study of the Danish 1936-2004 cleft twin cohort. *Am J Med Genet A*. 2010;152A:2468–2474.

[7] Grosen D, Bille C, Petersen I, Skytthe A, Hjelmborg JvB, Pedersen JK, et al. Risk of oral clefts in twins. *Epidemiology*. 2011;22:313–319.

[8] Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*. 2014;15:335–346.

[9] Breslow NE, Day NE. *Statistical Methods in Cancer Research. Volume 1 — The Analysis of Case-Control Studies*. Lyon: IARC Scientific Publications; 1980.

[10] Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA*. 2008;299:1335–1344.

[11] Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet*. 1987;51:227–233.

[12] Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*. 1993;53:1114–1126.

[13] Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. *Ann Hum Genet.* 2018;82:60–73.

[14] Gjerdevik M, Jugessur A, Haaland ØA, Romanowska J, Lie RT, Cordell HJ, et al. Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls. *BMC Bioinformatics.* 2019;20:165.

[15] Gjerdevik M, Gjessing HK, Romanowska J, Haaland ØA, Jugessur A, Czajkowski NO, et al. Design efficiency in genetic association studies. *Stat Med.* 2020; Epub ahead of print: DOI: 10.1002/sim.8476.

[16] Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Ann Hum Genet.* 2006;70:382–396.

[17] Gjessing HK. Haplin: analyzing case-parent triad and/or case-control data with SNP haplotypes; 2019. R package version 7.1.0. Available from: `https://people.uib.no/gjessing/genetics/software/haplin`.

[18] Ziegler A, König IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. 2nd ed. Weinheim: Wiley-Blackwell; 2012.

[19] Elston RC, Satagopan JM, Sun S. Genetic terminology. *Methods Mol Biol.* 2012;850:1–9.

[20] Wikimedia Commons. File:Dna-SNP.svg — Wikimedia Commons, the free media repository; 2017. [Online; accessed 3-September-2019]. Available from: `https://commons.wikimedia.org/w/index.php?title=File:Dna-SNP.svg&oldid=230242693`.

[21] Chial H. Mendelian genetics: patterns of inheritance and single-gene disorders. *Nature Education.* 2008;1:63.

[22] Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature.* 1983;306:234–238.

[23] White R, Woodward S, Leppert M, O'Connell P, Hoff M, Herbst J, et al. A closely linked genetic marker for cystic fibrosis. *Nature.* 1985;318:382–384.

[24] Jugessur A, Skare Ø, Harris JR, Lie RT, Gjessing HK. Using offspring-parent triads to study complex traits: a tutorial based on orofacial clefts. *Norsk Epidemiologi.* 2012;21:251–267.

[25] van Vliet J, Oates NA, Whitelaw E. Epigenetic mechanisms in the context of complex diseases. *Cell Mol Life Sci.* 2007;64:1531–1538.

[26] Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. *Nat Rev Genet.* 2013;14:609–617.

[27] Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* 1996;273:1516–1517.

[28] Risch NJ. Searching for genetic determinants in the new millennium. *Nature.* 2000;405:847–856.

[29] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 2008;9:356–369.

[30] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med.* 2002;4:45–61.

[31] Ioannidis JPA. Genetic associations: false or true? *Trends Mol Med.* 2003;9:135–138.

[32] Hutchison KE, Stallings M, McGeary J, Bryan A. Population stratification in the candidate gene study: fatal threat or red herring? *Psychol Bull.* 2004;130:66–79.

[33] Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet.* 2008;40:17–22.

[34] Mossey P, Castillia E. Global registry and database on craniofacial anomalies. Geneva: World Health Organization; 2003.

[35] Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, et al. Disruption of an AP-2$\alpha$ binding site in an *IRF6* enhancer is strongly associated with cleft lip. *Nat Genet.* 2008;40:1341–1347.

[36] Stanier P, Moore GE. Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts. *Hum Mol Genet.* 2004;13:R73–R81.

[37] Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *Lancet.* 2009;374:1773–1785.

[38] Jugessur A, Murray JC. Orofacial clefting: recent insights into a complex trait. *Curr Opin Genet Dev.* 2005;15:270–278.

[39] Buyske S. Maternal genotype effects can alias case genotype effects in case-control studies. *Eur J Hum Genet.* 2008;16:783–785.

[40] Shi M, Christensen K, Weinberg CR, Romitti P, Bathum L, Lozada A, et al. Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *Am J Hum Genet.* 2007;80:76–90.

[41] McGinnis R, Steinthorsdottir V, Williams NO, Thorleifsson G, Shooter S, Hjartardottir S, et al. Variants in the fetal genome near *FLT1* are associated with risk of preeclampsia. *Nat Genet.* 2017;49:1255–1260.

[42] Wang MH, Cordell HJ, Van Steen K. Statistical methods for genome-wide association studies. *Semin Cancer Biol.* 2019;55:53–60.

[43] Bartolomei MS, Tilghman SM. Genomic imprinting in mammals. *Annu Rev Genet.* 1997;31:493–525.

[44] Reik W, Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet.* 2001;2:21–32.

[45] Bartolomei MS. Genomic imprinting: employing and avoiding epigenetic processes. *Genes Dev.* 2009;23:2124–2133.

[46] Butler MG. Imprinting disorders: non-Mendelian mechanisms affecting growth. *J Pediatr Endocrinol Metab.* 2002;15:1279–1288.

[47] Wang KS, Liu X, Zhang Q, Aragam N, Pan Y. Parent-of-origin effects of *FAS* and *PDLIM1* in attention-deficit/hyperactivity disorder. *J Psychiatry Neurosci.* 2012;37:46–52.

[48] Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, et al. Parental origin of sequence variants associated with complex diseases. *Nature.* 2009;462:868–874.

[49] Lyssenko V, Groop L, Prasad RB. Genetics of type 2 diabetes: it matters from which parent we inherit the risk. *Rev Diabet Stud.* 2015;12:233–242.

[50] Connolly S, Heron EA. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Brief Bioinform.* 2015;16:429–448.

[51] Guilmatre A, Sharp AJ. Parent of origin effects. *Clin Genet.* 2012;81:201–209.

[52] Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol.* 1998;148:893–901.

[53] Barker DJP, Winter PD, Osmond C, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. *Lancet.* 1989;2:577–580.

[54] Kajantie E, Osmond C, Barker DJP, Forsén T, Phillips DIW, Eriksson JG. Size at birth as a predictor of mortality in adulthood: a follow-up of 350 000 person-years. *Int J Epidemiol.* 2005;34:655–663.

[55] Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson BJ, Young AI, Thorgeirsson TE, et al. The nature of nurture: effects of parental genotypes. *Science.* 2018;359:424–428.

[56] Hager R, Cheverud JM, Wolf JB. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics.* 2008;178:1755–1762.

[57] Sinsheimer JS, Palmer CGS, Woodward JA. Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. *Genet Epidemiol.* 2003;24:1–13.

[58] Haaland ØA, Lie RT, Romanowska J, Gjerdevik M, Gjessing HK, Jugessur A. A genome-wide search for gene-environment effects in isolated cleft lip with or without cleft palate triads points to an interaction between maternal periconceptional vitamin use and variants in *ESRRG*. *Front Genet.* 2018;9:60.

[59] Haaland ØA, Jugessur A, Gjerdevik M, Romanowska J, Shi M, Beaty TH, et al. Genome-wide analysis of parent-of-origin interaction effects with environmental exposure (PoOxE): an application to European and Asian cleft palate trios. *PLoS One.* 2017;12:e0184358.

[60] Haaland ØA, Romanowska J, Gjerdevik M, Lie RT, Gjessing HK, Jugessur A. A genome-wide scan of cleft lip triads identifies parent-of-origin interaction effects between *ANK3* and maternal smoking, and between *ARHGEF10* and alcohol consumption. *F1000Res.* 2019;8:960 (Version 2).

[61] Romanowska J, Haaland ØA, Jugessur A, Gjerdevik M, Xu Z, Taylor J, et al. Gene-methylation interactions: discovering region-wise DNA methylation levels that modify SNP-associated disease risk. *bioRxiv.* 2019;593053 [preprint].

[62] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet.* 2012;13:484–492.

[63] Ho GYF, Bailey-Wilson JE. The transmission/disequilibrium test for linkage on the X chromosome. *Am J Hum Genet.* 2000;66:1158–1160.

[64] Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered.* 2008;66:87–98.

[65] Zhang L, Martin ER, Chung RH, Li YJ, Morris RW. X-LRT: a likelihood approach to estimate genetic risks and test association with X-linked markers using a case-parents design. *Genet Epidemiol.* 2008;32:370–380.

[66] Jugessur A, Skare Ø, Lie RT, Wilcox AJ, Christensen K, Christiansen L, et al. X-linked genes and risk of orofacial clefts: evidence from two population-based studies in Scandinavia. *PLoS One.* 2012;7:e39240.

[67] Skare Ø, Gjessing HK, Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, et al. A new approach to chromosome-wide analysis of X-linked markers identifies new associations in Asian and European case-parent triads of orofacial clefts. *PLoS One.* 2017;12:e0183772.

[68] Sharma R, Harris VM, Cavett J, Kurien BT, Liu K, Koelsch KA, et al. Rare X chromosome abnormalities in systemic lupus erythematosus and Sjögren's syndrome. *Arthritis Rheumatol.* 2017;69:2187–2192.

[69] Skare Ø, Lie RT, Haaland ØA, Gjerdevik M, Romanowska J, Gjessing HK, et al. Analysis of parent-of-origin effects on the X chromosome in Asian and European orofacial cleft triads identifies associations with *DMD*, *FGF13*, *EGFL6*, and additional loci at Xp22.2. *Front Genet.* 2018;9:25.

[70] Liu H, Li S, Wang X, Zhu J, Wei Y, Wang Y, et al. DNA methylation dynamics: identification and functional annotation. *Brief Funct Genomics.* 2016;15:470–484.

[71] Cordell HJ, Clayton DG. Genetic association studies. *Lancet.* 2005;366:1121–1131.

[72] Clayton D, Hills M. *Statistical Models in Epidemiology.* Oxford: Oxford University Press; 1993.

[73] Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. *Curr Protoc Hum Genet.* 2017;95:1.22.1–1.22.23.

[74] Self SG, Longton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics.* 1991;47:53–61.

[75] Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet.* 1993;52:1085–1093.

[76] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993;52:506–516.

[77] Schaid DJ, Sommer SS. Comparison of statistics for candidate-gene association studies using cases and parents. *Am J Hum Genet.* 1994;55:402–409.

[78] Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol.* 1996;13:423–449.

[79] Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet.* 2000;66:251–261.

[80] Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet.* 2005;77:627–636.

[81] Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet.* 2004;12:964–970.

[82] Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet.* 2005;76:592–608.

[83] Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. *Am J Epidemiol.* 2008;168:541–547.

[84] Vermeulen SH, Shi M, Weinberg CR, Umbach DM. A hybrid design: case-parent triads supplemented by control-mother dyads. *Genet Epidemiol.* 2009;33:136–144.

[85] Skare Ø, Jugessur A, Lie RT, Wilcox AJ, Murray JC, Lunde A, et al. Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. *Ann Hum Genet.* 2012;76:221–236.

[86] Stewart WCL, Cerise J. Increasing the power of association studies with affected families, unrelated cases and controls. *Front Genet.* 2013;4:200.

[87] Infante-Rivard C, Mirea L, Bull SB. Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am J Epidemiol.* 2009;170:657–664.

[88] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans R Soc Lond A.* 1933;231:289–337.

[89] Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol.* 2006;164:609–614.

[90] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet.* 2007;39:17–23.

[91] Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* 2007;17:1520–1528.

[92] Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics.* 2003;19:149–150.

[93] Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med.* 2014;33:1946–1978.

[94] van der Vaart AW. *Asymptotic Statistics*. Cambridge: Cambridge University Press; 2000.

[95] Noether GE. On a theorem of Pitman. *Ann Math Stat.* 1955;26:64–68.

[96] Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet.* 1995;56:811–812.

[97] Sun F, Flanders WD, Yang Q, Khoury MJ. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. *Am J Epidemiol.* 1999;150:97–104.

[98] Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet.* 1999;65:229–235.

[99] Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet.* 1998;62:969–978.

[100] Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet.* 2002;70:124–141.

[101] Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol.* 2004;26:167–185.

[102] Cordell HJ. Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genet Epidemiol.* 2004;26:186–205.

[103] Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet.* 1999;64:1186–1193.

[104] Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol.* 2011;35:19–45.

[105] Howey R, Cordell HJ. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinformatics.* 2012;13:149.

[106] Howey R, Mamasoula C, Töpf A, Nudel R, Goodship JA, Keavney BD, et al. Increased power for detection of parent-of-origin effects via the use of haplotype estimation. *Am J Hum Genet.* 2015;97:419–434.

[107] R Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2019. Available from: `https://www.R-project.org/`.

[108] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B.* 1977;39:1–38.

[109] Shi M, Umbach DM, Weinberg CR. Testing haplotype-environment interactions using case-parent triads. *Hum Hered.* 2010;70:23–33.

[110] Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley; 2013.

[111] Wang M, Xu S. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity.* 2019;123:287–306.

[112] Schaid DJ. Disease-marker association. In: Elston R, Olson J, Palmer L, editors. Biostatistical Genetics and Genetic Epidemiology. Wiley reference series in biostatistics. West Sussex, UK: Wiley; 2002. p. 206–217.

[113] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559–575.

[114] Baker SG. The multinomial-poisson transformation. *J R Stat Soc Ser D.* 1994;43:495–504.

[115] Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near *MAFB* and *ABCA4*. *Nat Genet.* 2010;42:525–529.

[116] Beaty TH, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, et al. Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genet Epidemiol.* 2011;35:469–478.

[117] Shi M, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, et al. Genome wide study of maternal and parent-of-origin effects on the etiology of orofacial clefts. *Am J Med Genet A.* 2012;158A:784–794.

[118] Dudbridge F. UNPHASED. Version 3.1.7. London School of Hygiene and Tropical Medicine; June, 2013.

[119] Wise AS, Shi M, Weinberg CR. Family-based multi-SNP X chromosome analysis using parent information. *Front Genet.* 2016;7:20.

[120] Lange K, Sinsheimer JS, Sobel E. Association testing with Mendel. *Genet Epidemiol.* 2005;29:36–50.

[121] Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. Mendel: the Swiss army knife of genetic analysis programs. *Bioinformatics.* 2013;29:1568–1570.

[122] Vermunt JK. *LEM: A General Program for the Analysis of Categorical Data*. Tilburg University; 1997.

[123] Vermunt JK. *Log-Linear Models for Event Histories*. Thousand Oaks, CA: Sage Publications; 1997.

[124] van Den Oord EJCG, Vermunt JK. Testing for linkage disequilibrium, maternal effects, and imprinting with (in)complete case-parent triads, by use of the computer program LEM. *Am J Hum Genet.* 2000;66:335–338.

[125] Clayton D. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet.* 1999;65:1170–1177.

[126] Shi M, Umbach DM, Weinberg CR. Identification of risk-related haplotypes with the use of multiple SNPs from nuclear families. *Am J Hum Genet.* 2007;81:53–66.

[127] Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, Weinberg CR, et al. Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PLoS One.* 2009;4:e5385.

[128] Schwender H, Li Q, Neumann C, Taub MA, Younkin SG, Berger P, et al. Dectecting disease variants in case-parent trio studies using the Bioconductor software package trio. *Genet Epidemiol.* 2014;38:516–522.

[129] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics.* 2007;23:1294–1296.

[130] Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med.* 2002;21:35–50.

[131] Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol.* 2002;155:478–484.

[132] Gauderman WJ. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol.* 2003;25:327–338.

[133] Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: tools for family-based association studies. *Am J Hum Genet.* 2004;74:367–369.

[134] Van Steen K, Lange C. PBAT: a comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum Genomics.* 2005;2:67–69.

[135] Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet.* 2006;7:385–394.

[136] Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. *Genet Epidemiol.* 2000;19:S36–S42.

[137] Lange C, DeMeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet.* 2002;71:1330–1341.

[138] Lange C, Laird NM. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet Epidemiol.* 2002;23:165–180.

[139] Lange C, Laird NM. Power calculations for a general class of family-based association tests: dichotomous traits. *Am J Hum Genet.* 2002;71:575–584.

[140] Weinberg CR, Shi M. The genetics of preterm birth: using what we know to design better association studies. *Am J Epidemiol.* 2009;170:1373–1381.

[141] Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ.* 2006;332:1080.

[142] Thomas DC. Case-parents design for gene-environment interaction by Schaid. *Genet Epidemiol.* 2000;19:461–463.

[143] Shen YC, Fan JH, Edenberg HJ, Li TK, Cui YH, Wang YF, et al. Polymorphism of *ADH* and *ALDH* genes among four ethnic groups in China and effects upon the risk for alcoholism. *Alcohol Clin Exp Res.* 1997;21:1272–1277.

[144] Wahlund S. Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas.* 1928;11:65–106.

[145] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57:289–300.

[146] Jung SH. Sample size for FDR-control in microarray data analysis. *Bioinformatics.* 2005;21:3097–3104.

[147] Wang S, Yu Z, Miller RL, Tang D, Perera FP. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Hum Hered.* 2011;71:196–208.

[148] Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 2011;12:R10.

[149] Hedman ÅK, Zilmer M, Sundström J, Lind L, Ingelsson E. DNA methylation patterns associated with oxidative stress in an ageing population. *BMC Med Genomics.* 2016;9:72.

[150] Gao X, Thomsen H, Zhang Y, Breitling LP, Brenner H. The impact of methylation quantitative trait loci (mQTLs) on active smoking-related DNA methylation changes. *Clin Epigenetics.* 2017;9:87.

[151] Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat Rev Genet.* 2019;20:235–248.

[152] Schaid DJ, Rowland C. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am J Hum Genet.* 1998;63:1492–1506.

[153] Martin ER, Bass MP, Hauser ER, Kaplan NL. Accounting for linkage in family-based tests of association with missing parental genotypes. *Am J Hum Genet.* 2003;73:1016–1026.

[154] Shi M, Umbach DM, Weinberg CR. Case-sibling studies that acknowledge unstudied parents and permit the inclusion of unmatched individuals. *Int J Epidemiol.* 2013;42:298–307.

[155] Curtis D. Use of siblings as controls in case-control association studies. *Ann Hum Genet.* 1997;61:319–333.

[156] Wright ML, Dozmorov MG, Wolen AR, Jackson-Cook C, Starkweather AR, Lyon DE, et al. Establishing an analytic pipeline for genome-wide DNA methylation. *Clin Epigenetics.* 2016;8:45.

[157] White CC, Yang HS, Yu L, Chibnik LB, Dawe RJ, Yang J, et al. Identification of genes associated with dissociation of cognitive performance and neuropathological burden: multistep analysis of genetic, epigenetic, and transcriptional data. *PLoS Med.* 2017;14:e1002287.

[158] Shilpi A, Bi Y, Jung S, Patra SK, Davuluri RV. Identification of genetic and epigenetic variants associated with breast cancer prognosis by integrative bioinformatics analysis. *Cancer Inform.* 2017;16:1–13.

[159] Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet.* 2018;19:129–147.

[160] Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19:491–504.

[161] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* 2018;562:203–209.

[162] Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort profile update: the Norwegian Mother and Child Cohort study (MoBa). *Int J Epidemiol.* 2016;45:382–388.

[163] Helgeland Ø, Vaudel M, Juliusson PB, Lingaas Holmen O, Juodakis J, Bacelis J, et al. Genome-wide association study reveals dynamic role of genetic variation in infant and early childhood growth. *Nat Commun.* 2019;10:4448.

[164] Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. *Nature.* 2017;542:186–190.

[165] Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet.* 2014;95:5–23.

[166] Wainschtein P, Jain DP, Yengo L, Zheng Z, TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv.* 2019;588020 [preprint].

[167] Young AI. Solving the missing heritability problem. *PLoS Genet.* 2019;15:e1008222.

[168] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* 2012;90:7–24.

[169] Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics.* 2015;31:1466–1468.

[170] Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* 2013;9:e1003348.

[171] Barcellos SH, Carvalho LS, Turley P. Education can reduce health differences related to genetic risk of obesity. *Proc Natl Acad Sci USA.* 2018;115:E9765–E9772.

[172] Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* 2018;50:1219–1224.

[173] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet.* 2018;19:581–590.

[174] Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans. *Genetics.* 2019;211:1131–1141.

[175] Young AI, Benonisdottir S, Przeworski M, Kong A. Deconstructing the sources of genotype-phenotype associations in humans. *Science.* 2019;365:1396–1400.

[176] Chen R, Snyder M. Promise of personalized omics to precision medicine. *Wiley Interdiscip Rev Syst Biol Med.* 2013;5:73–82.

[177] Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci.* 2015;282:20151684.

[178] Marquart J, Chen EY, Prasad V. Estimation of the percentage of US patients with cancer who benefit from genome-driven oncology. *JAMA Oncol.* 2018;4:1093–1098.

[179] Joyner MJ, Paneth N. Promises, promises, and precision medicine. *J Clin Invest.* 2019;129:946–948.

# Appendices

## Appendix I—Haplin commands for the example in Section 1.9.1

The Haplin commands below were used to simulate and analyze the data in Section 1.9.1. However, small adjustments have been made for renaming the SNP and alleles. Note that small discrepancies may occur depending on the processor architecture and operating system. Please consult the **R** help files for a description of the Haplin functions and their arguments.

```
## Load Haplin
library(Haplin)

## Set seed
set.seed(1231)

## Simulate data in Haplin format using 340 case-parent triads and
## 460 control-parent triads, a MAF of 0.1, and a relative risk of 1.6

hapSim(nall = c(2), n.strata = 1, cases = c(mfc = 340), controls = c(mfc = 460),
        haplo.freq = c(0.9, 0.1), RR = c(1, 1.6), RRstar = c(1, 1),
        n.sim = 1, dire = "haplinData")

## Read and prepare data for analysis

data <- genDataRead(file.in = "haplinData/sim1.dat",
        file.out = "haplinData", dir.out = "haplinData",
        format = "haplin", n.vars = 1, allele.sep = " ", col.sep = " ")

prep.data <- genDataPreprocess(data.in = data, design = "cc.triad",
        file.out = "prep_data", dir.out = "haplinData")

## Run Haplin analysis
res <- haplin(prep.data, response = "mult", design = "cc.triad",
        ccvar = 1, reference = "ref.cat")

## Get full output
haptable(res)

## Plot results
plot(res, filename = "haplin_run.png")
```

# Appendix II—Tables 1 and 2 from Paper I

The layout of Tables 1 and 2 in the published paper makes them somewhat difficult to read. To better illustrate the classification of genetic effects, the submitted versions are also attached.

Table 1, Paper I: PoO, GxE and PoOxE effects for cleft-palate-only example SNPs

**a) rs7516430, *CHD1L*** [1]

| Test effect | Stratum | $\mathrm{RR}_M$ | $\mathrm{RR}_F$ | $\mathrm{RR}_M/\mathrm{RR}_F$ |
|---|---|---|---|---|
| | $\mathrm{RR}_\mathrm{S}$ | 1.79 | 0.52 | 3.42 (1.86, 6.15) |
| PoO effects* | $\mathrm{RR}_\mathrm{NS}$ | 1.79 | 0.52 | 3.42 (1.86, 6.15) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 1 (-) | 1 (-) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 1.22 | 1.22 | 1 (-) |
| GxE effects** | $\mathrm{RR}_\mathrm{NS}$ | 1.06 | 1.06 | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 1.15 (0.51, 2.61) | 1.15 (0.51, 2.61) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 1.88 | 0.66 | 2.83 (0.90, 8.63) |
| PoOxE effects | $\mathrm{RR}_\mathrm{NS}$ | 1.76 | 0.48 | 3.68 (1.80, 7.37) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 1.07 (0.43, 2.69) | 1.40 (0.40, 4.83) | 0.77 (0.20, 2.91) |

**b) r470563, *ZNF236*** [2]

| Test effect | Stratum | $\mathrm{RR}_M$ | $\mathrm{RR}_F$ | $\mathrm{RR}_M/\mathrm{RR}_F$ |
|---|---|---|---|---|
| | $\mathrm{RR}_\mathrm{S}$ | 0.95 | 1.07 | 0.89 (0.67, 1.17) |
| PoO effects* | $\mathrm{RR}_\mathrm{NS}$ | 0.95 | 1.07 | 0.89 (0.67, 1.17) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 1 (-) | 1 (-) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 0.48 | 0.48 | 1 (-) |
| GxE effects** | $\mathrm{RR}_\mathrm{NS}$ | 1.15 | 1.15 | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 0.42 (0.26, 0.68) | 0.42 (0.26, 0.68) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 0.44 | 0.52 | 0.86 (0.39, 1.87) |
| PoOxE effects | $\mathrm{RR}_\mathrm{NS}$ | 1.09 | 1.22 | 0.89 (0.66, 1.20) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 0.41 (0.21, 0.79) | 0.42 (0.23, 0.80) | 0.96 (0.41, 2.24) |

**c) rs2964137, *ICE1*** [3]

| Test effect | Stratum | $\mathrm{RR}_M$ | $\mathrm{RR}_F$ | $\mathrm{RR}_M/\mathrm{RR}_F$ |
|---|---|---|---|---|
| | $\mathrm{RR}_\mathrm{S}$ | 1.42 | 1.06 | 1.34 (0.90, 1.97) |
| PoO effects* | $\mathrm{RR}_\mathrm{NS}$ | 1.42 | 1.06 | 1.34 (0.90, 1.97) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 1 (-) | 1 (-) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 1.16 | 1.16 | 1 (-) |
| GxE effects** | $\mathrm{RR}_\mathrm{NS}$ | 1.25 | 1.25 | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 0.93 (0.54, 1.60) | 0.93 (0.54, 1.60) | 1 (-) |
| | $\mathrm{RR}_\mathrm{S}$ | 0.53 | 2.57 | 0.21 (0.09, 0.46) |
| PoOxE effects | $\mathrm{RR}_\mathrm{NS}$ | 1.88 | 0.85 | 2.22 (1.41, 3.43) |
| | $\mathrm{RR}_\mathrm{S}/\mathrm{RR}_\mathrm{NS}$ | 0.28 (0.13, 0.58) | 3.03 (1.45, 6.35) | 0.09 (0.04, 0.24) |

* PoO effects were estimated without stratifying on exposure. The rows corresponding to environmental strata are therefore equal by assumption.

** GxE effects were estimated without stratifying on parental origin. The columns related to $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are therefore equal by assumption.

- The estimates are relative to the most frequent allele
- $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are the relative risks depending on parental origin
- $\mathrm{RR}_\mathrm{NS}$ and $\mathrm{RR}_S$ are the relative risks depending on exposure status (non-smokers or smokers)

[1] Overall allele frequencies: A 0.88; T 0.12; Europeans only

[2] Overall allele frequencies: C 0.57; G 0.43; Whole sample

[3] Overall allele frequencies: G 0.52; C 0.48; Europeans only

Table 2, Paper I: PoOxE effects for cleft-palate-only example haplotypes

rs2964447-rs2964137-rs6868526, *ICE1*

| Haplotype | Stratum | $\mathrm{RR}_M$ | $\mathrm{RR}_F$ | $\mathrm{RR}_M/\mathrm{RR}_F$ |
|---|---|---|---|---|
| T-G-C | $\mathrm{RR}_{\mathrm{S}}$ | 1.99 | 0.49 | 4.04 (1.75, 9.25) |
| | $\mathrm{RR}_{\mathrm{NS}}$ | 0.52 | 1.04 | 0.50 (0.31, 0.82) |
| | $\mathrm{RR}_{\mathrm{S}}/\mathrm{RR}_{\mathrm{NS}}$ | 3.79 (1.74, 8.22) | 0.47 (0.21, 1.05) | 7.98 (3.07, 20.77) |
| T-G-G | $\mathrm{RR}_{\mathrm{S}}$ | 1.30 | 0.24 | 5.35 (1.51, 18.19) |
| | $\mathrm{RR}_{\mathrm{NS}}$ | 0.68 | 1.30 | 0.52 (0.29, 0.96) |
| | $\mathrm{RR}_{\mathrm{S}}/\mathrm{RR}_{\mathrm{NS}}$ | 1.89 (0.70, 5.07) | 0.19 (0.06, 0.62) | 10.13 (2.55, 40.19) |

- Reference haplotype: A-C-C
- Overall haplotype frequencies: A-C-C 0.48; T-G-C 0.36; T-G-G 0.16; Europeans only
- $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are the relative risks depending on parental origin
- $\mathrm{RR}_{\mathrm{NS}}$ and $\mathrm{RR}_S$ are the relative risks depending on exposure status (non-smokers or smokers)

Paper I

# Parent-of-origin-environment interactions in case-parent triads with or without independent controls

Received: 15 March 2017 | Accepted: 5 September 2017

DOI: 10.1111/ahg.12224

ORIGINAL ARTICLE

WILEY Annals of human genetics

# Parent-of-origin-environment interactions in case-parent triads with or without independent controls

Miriam Gjerdevik[1,2] | Øystein A. Haaland[1] | Julia Romanowska[1,3] | Rolv T. Lie[1,4] | Astanand Jugessur[1,2,5] | Håkon K. Gjessing[1,5]

[1]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[2]Department of Genetic Research and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway

[3]Compitional Biology Unit, University of Bergen, Bergen, Norway

[4]Department of Health Registries, Norwegian Institute of Public Health, Oslo, Norway

[5]Centre for Fertility and Health (CeFH), Norwegian Institute of Public Health, Oslo, Norway

**Correspondence**
Miriam Gjerdevik, Department of Global Public Health and Primary Care, University of Bergen, N-5020 Bergen, Department of Genetic Research and Bioinformatics, Norwegian Institute of Public Health, N-0403 Oslo, Norway.
E-mail: miriam.gjerdevik@uib.no

**Funding information**
This research was supported by the Bergen Medical Research Foundation (BMFS) grant 807191, by the Research Council of Norway (RCN)'s Biobank Norway grant 245464/F50, and by the RCN through its Centres of Excellence funding scheme, grant 262700.

**Abstract**

With case–parent triad data, one can frequently deduce parent of origin of the child's alleles. This allows a parent-of-origin (PoO) effect to be estimated as the ratio of relative risks associated with the alleles inherited from the mother and the father, respectively. A possible cause of PoO effects is DNA methylation, leading to genomic imprinting. Because environmental exposures may influence methylation patterns, gene–environment interaction studies should be extended to allow for interactions between PoO effects and environmental exposures (i.e., PoOxE). One should thus search for loci where the environmental exposure modifies the PoO effect.

We have developed an extensive framework to analyze PoOxE effects in genome-wide association studies (GWAS), based on complete or incomplete case–parent triads with or without independent control triads. The interaction approach is based on analyzing triads in each exposure stratum using maximum likelihood estimation in a log-linear model. Interactions are then tested applying a Wald-based posttest of parameters across strata. Our framework includes a complete setup for power calculations. We have implemented the models in the R software package Haplin.

To illustrate our PoOxE test, we applied the new methodology to top hits from our previous GWAS, assessing whether smoking during the periconceptional period modifies PoO effects on cleft palate only.

KEYWORDS

case–parent triad, gene–environment interaction, hybrid design, imprinting, parent-of-origin, power and sample size calculation, trios

## 1 | INTRODUCTION

A large number of human traits can be classified as complex, in the sense that they are assumed to be influenced by multiple genes and their interactions with environmental or behavioral factors (Pasaniuc & Price, 2016). Although thousands of genome-wide association studies (GWAS) have been conducted since the turn of the millennium, for most complex traits the genetic variants identified thus far explain only a small fraction of the phenotypic variation attributed to genetic effects (Manolio et al., 2009). This has underscored the need to investigate disease mechanisms beyond simple genetic effects alone. One example is gene–environment interactions (GxE), where the genetic effects are modified by environmental exposures. For instance, Shi et al. (2007) have shown that maternal cigarette smoking in the periconceptional period can modify the association between single nucleotide polymorphisms (SNPs) and orofacial clefts.

With access to case–parent triad data, where an offspring and his/her parents have been genotyped, other genetic effects such as parent-of-origin (PoO) effects can be assessed. A PoO effect refers to the situation where the effect of a particular allele in the child depends on whether it is inherited from the mother or the father (Lawson, Cheverud, & Wolf, 2013; Connolly & Heron, 2014). For example, an allele might be protective when inherited from the mother but detrimental when inherited from the father. One example of a PoO effect is genomic imprinting, an epigenetic phenomenon where one of the inherited parental alleles is expressed whereas the other is silenced (Bartolomei & Tilghman, 1997; Reik & Walter, 2001). Although PoO effects are often used interchangeably with imprinting (Lawson et al., 2013), we here define PoO effects in statistical terms to mean an interaction effect; a PoO effect occurs if the phenotypic risk varies according to the parental origin of the variant allele.

In recent years, a growing number of studies have aimed to identify PoO and GxE effects separately for a wide range of diseases. However, it is reasonable to assume that the combined interaction effect (PoOxE effect) may also play an important role in complex traits. In our context, this means that the observed PoO effect may vary across environmental strata, which is plausible from a biologic perspective. A known cause of imprinting is DNA methylation in the germline. It is possible that maternal environmental exposures influencing methylation patterns might also influence the effects of maternally and paternally inherited alleles in unequal measures.

Conceivably, PoOxE effects may appear in different ways. The allele in question might increase risk only when transmitted from exposed mothers. A PoOxE effect may also be observed if the allele is protective to the child only when inherited from unexposed mothers but with no particular effect in the other situations. In principle, there might even be a “qualitative” interaction where the genetic effect is reversed. For instance, an allele might increase risk when inherited from exposed mothers and decrease risk when inherited from unexposed mothers, and concurrently decrease risk when inherited from exposed fathers and increase risk when inherited from unexposed fathers.

Another factor that needs to be controlled for in PoOxE models is the possible presence of maternal genetic effects. Maternal genetic effects occur when the genotype of the mother affects the phenotype of the child, regardless of the genetic material that has been transferred from mother to child (Connolly & Heron, 2014). Alleles carried by the mother may influence fetal development directly, for example, through maternal metabolic factors (Guilmatre & Sharp, 2012). This effect is distinct from PoO effects, in which we compare the effect of alleles *in the child*, depending on whether they were inherited from the mother or the father (Howey et al., 2015). Maternal genetic effects must therefore be estimated primarily from the nontransmitted allele of the mother, and appropriate models for PoOxE effects should allow maternal and PoO effects to be estimated simultaneously. Clearly, maternal effects are particularly important to studies of perinatal disorders.

Wang, Yu, Miller, Tang, and Perera (2011) previously introduced a test to screen for interactions between imprinted genes and environmental exposures. Still, there is a need to develop more general methods to investigate the joint effects of PoO and GxE (Lawson et al., 2013, p. 616). To address this gap in knowledge, we propose a novel approach that enables a full investigation of PoOxE effects. We develop our model for PoOxE within a flexible maximum-likelihood framework based on log-linear models (Gjessing & Lie, 2006; Skare et al., 2012; Jugessur, Skare, Harris, Lie, & Gjessing, 2012a), originally described in Wilcox, Weinberg, and Lie (1998), Weinberg, Wilcox, and Lie (1998), and Gjessing and Lie (2006). Our main study unit is the case-parent triad, but it can be extended to include independent control children or control triads in a hybrid design (Weinberg & Umbach, 2005). Note that control triads are optional because the nontransmitted parental alleles implicitly serve as pseudocontrols (Knapp, Seuchter, & Baur, 1993; Schaid & Sommer, 1993; Cordell, Barratt, & Clayton, 2004; Cordell, 2004). Moreover, we use an expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to accommodate missing parents in mother–offspring or father–offspring dyads. A full implementation of our models is provided in Haplin, a flexible R package for genetic association analyses of single SNPs or haplotypes (Gjessing & Lie, 2006). The implementation uses parallel processing of SNPs, which makes GWAS analyses feasible. Haplin performs both testing and estimation of genetic effects. The framework also incorporates analyses of X-chromosome SNPs in a natural way.

In statistical terms, PoO analyses are interaction analyses; the effect of an allele in the child may be modified by its parent of origin. In contrast, regular fetal-effect analyses assume that the effect of an allele in the child is independent of whether it is transmitted from the mother or the father, that is, the effect is estimated without stratifying on parental origin. Higher sample sizes are thus required for PoO analyses to achieve the same statistical power as in regular fetal-effect analyses. Accordingly, PoOxE analyses can be seen as second-order interaction analyses. Hence, an even larger sample size is needed for a PoOxE analysis than for the corresponding PoO or GxE analysis to obtain the same statistical power. We therefore provide a thorough discussion of the power for PoOxE analyses and provide software to compute power for all relevant scenarios.

The article is structured as follows. In the Methods section, we first provide relevant background information and present the sampling and penetrance models. Next, we introduce our PoOxE test and derive the statistical methodology for single-SNP analysis, and we also explain how PoOxE analyses can be carried out for SNPs on the X-chromosome. We conclude the Methods section by presenting a previously published case triad study of orofacial clefts. In the Results section, we illustrate our PoOxE approach by using Haplin to analyze genetic triad data from the cleft study. We then assess the operating characteristics of the PoOxE test by investigating its power and attained significance level. The appendix includes a detailed discussion of PoOxE effects for haplotypes (Appendix A.1). Additionally, issues pertaining to sample size and power calculation are considered, and we present formulae and algorithms for our power computations (Appendix A.2). Haplin commands for estimating PoO, GxE and PoOxE effects on candidate genes are provided in the Supporting Information (S1). Statistical power calculations in Haplin are also covered in detail.

## 2 | METHODS

### 2.1 | Sampling and penetrance model

The likelihood model is based on a log-linear model for the observed triad frequencies, conditional on the child being a case. Optionally, independent controls or control triads can be added to improve estimation of allele/haplotype frequencies. In this section, we describe the underlying sampling and penetrance model. A more detailed derivation of the log-linear model is provided elsewhere (Gjessing & Lie, 2006).

We consider a single, multi-allelic locus with $K$ alleles $A_1, A_2, \dots, A_K$, with corresponding population allele frequencies $p_1, p_2, \dots, p_K$. The genotypes for the mother, father, and child are denoted by $M$, $F$, and $C$, respectively, and the full triad as $(M, F, C) = (A_iA_j,\ A_kA_l,\ A_jA_l)$. For notational convenience, we assume that the second allele from the mother and the second allele from the father are transmitted to the child; that is, the full triad $(M, F, C)$ can thus be described by the mating type $(M,\ F) = (A_iA_j,\ A_kA_l)$.

The sampling model should describe the distribution of $(M, F, C)$, conditional on the child being a case. If $D$ denotes the event that the child is a case, Bayes' theorem allows our sampling model to be written as

$$P(M, F, C|D) = P(D|M, F, C)P(M, F, C)/P(D). \quad (1)$$

The disease prevalence, $P(D)$, cannot be observed directly from the case triad distribution and serves as a normalizing constant only. Assuming a population in Hardy–Weinberg equilibrium (HWE) with random mating and Mendelian transmission, we have

$$P(M, F, C) = P(A_iA_j, A_kA_l) = p_ip_jp_kp_l.$$

Although the HWE assumption can be avoided using a more detailed parameterization (Weinberg et al., 1998; Gjessing & Lie, 2006), its inclusion in the model is convenient for computational efficiency and useful for reconstructing haplotypes. However, analyses should always include a strategy for checking large deviations from HWE because such deviations may be indicative of data issues. Top hits from a GWAS analysis should always be further investigated; Haplin performs a test for HWE on all SNPs.

The penetrance model, $P(D|M, F, C)$, describes the probability of a child having the disease, conditional on the triad genotype. Assigning different effects to the alleles depending on parental origin, a penetrance model for PoO effects is

$$P(D|A_iA_j, A_kA_l) = B{\cdot}\mathrm{RR}_{M,j}\mathrm{RR}_{F,l}\mathrm{RR}^*_{jl},$$

where $\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ are the risk increase (or decrease) associated with allele $A_j$, relative to the baseline risk level $B$, depending on whether the allele is transmitted from the mother or the father. The fraction $\mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$ is then a measure of the extent of the risk associated with allele $A_j$, depending on parental origin. The parameter $\mathrm{RR}^*_{jl}$ is included to allow homozygous individuals to have a risk that deviates from what would be expected from a multiplicative model (e.g., dominant or recessive patterns). To incorporate this deviation, we have that $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and that $\mathrm{RR}^*_{jl} = 1$ when $j \neq l$. Thus, if $\mathrm{RR}^*_j = 1$ for all $j$, the penetrance model is purely multiplicative. Note that $B$ is typically associated with the reference allele and functions only as a normalizing constant. Moreover, this model also applies to multi-allelic markers. The full sampling model (1) can then be parameterized as

$$\begin{aligned}P(M, F, C|D) &= P(A_iA_j, A_kA_l|D)\\ &= p_ip_jp_kp_l \cdot B{\cdot}\mathrm{RR}_{M,j}\mathrm{RR}_{F,l}\mathrm{RR}^*_{jl}/P(D).\end{aligned}$$

Conditional on the child being a case, the triad type frequencies follow a multinomial distribution, and the parameters from the relevant sampling model are readily estimated by the method of maximum likelihood. The EM algorithm can be used to accommodate missing information, including reconstructing unknown haplotype phase from multiple markers. To ensure that the model is not overparameterized, one commonly sets $RR = 1$ for a reference allele. Alternatively, population or reciprocal references can be used (Gjessing & Lie, 2006). Notice that throughout this article we assume a multiplicative dose–response relationship.

An important feature of the log-linear model is the possibility to incorporate and adjust for maternal effects. Specifically, PoO and maternal genetic effects can be addressed simultaneously by the model

$$P(D|A_iA_j, A_kA_l) = B \cdot \mathrm{RR}_{M,j}\mathrm{RR}_{F,l}\mathrm{RR}^*_{jl} \times \mathrm{RR}^{(M)}_i\mathrm{RR}^{(M)}_j\mathrm{RR}^{(M)*}_{ij},$$

where $\mathrm{RR}^{(M)}_i$ is the relative risk associated with allele $A_i$ carried by the mother, and $\mathrm{RR}^{(M)*}_{ij}$ is interpreted analogously to $\mathrm{RR}^*_{ij}$. We thus assume that the maternal alleles have a multiplicative effect on top of the fetal alleles. Note specifically that in a combined model, the PoO effect is estimated essentially by contrasting allele frequencies of transmitted alleles, depending on parental origin, whereas the maternal effect is estimated by contrasting the frequencies of nontransmitted alleles in case mothers with that of nontransmitted alleles in case fathers.

Note that the PoO model requires information on parental origin, which is not available for ambiguous (uninformative) triads. However, the EM algorithm is implemented in our software and uses maximum likelihood to account for unknown parental origin in ambiguous triads. Additionally, it will account for missing information on individuals, such as when some triads are reduced to mother–child dyads due to missing data on the father. The basic model relates to a single multi-allelic locus. In combination with the EM algorithm it extends directly to haplotypes over multiple loci by statistically reconstructing unknown haplotype phase (Gjessing & Lie, 2006).

## 2.2 | Parent-of-origin-environment interactions

Our PoOxE approach seamlessly integrates the PoO model with that of GxE. We therefore start by presenting and interpreting the PoO and GxE analyses separately, before combining them in the PoOxE test. The theory for PoOxE is here derived for a single SNP, but the extension to haplotypes is provided in Appendix A.1. We conclude the section by illustrating how PoOxE effects can be assessed on the X-chromosome. Relevant Haplin commands for investigating PoO, GxE, and PoOxE effects are provided in S1.

For a single SNP, let $\mathrm{RR}_M$ and $\mathrm{RR}_F$ denote the relative risks associated with the variant allele (i.e., the nonreference allele) if it is inherited from the mother or from the father, respectively. We define the PoO effect as the relative risk ratio $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F$. This fraction is a measure of the magnitude of the risk associated with the allele under study, depending on whether it is maternally or paternally derived. A ratio larger than one indicates a higher risk when the variant allele is inherited from the mother versus the father. If it is equal to 1, the variant allele increases (or decreases) the risk by the same amount regardless of parental origin, and there is no PoO effect. For instance, if the variant allele doubles the risk of disease independently of parental origin, this is a standard fetal association; as such, it would have been identified in a traditional search for fetal gene effects. Note that one can assume a priori that, for instance, the paternal allele has no effect (i.e., $\mathrm{RR}_F = 1$) and try to detect a "pure" imprinting effect $\mathrm{RR}_M$. This effect is, however, confounded with a standard fetal effect whenever the assumption $\mathrm{RR}_F = 1$ does not hold. Accordingly, we prefer to define our PoO test as a contrast between maternally and paternally derived allele risks.

Under the weak assumption of independence between exposure and child genotype conditional on parental mating type (Shi, Umbach, & Weinberg, 2010), interactions between genes and a categorical exposure variable can be incorporated into the log-linear framework. Our GxE analyses fit the log-linear model separately in each exposure stratum and consequently do not assume that allele frequencies are constant across strata. The model uses a Wald test to detect whether the relative risk estimates differ significantly across the exposure levels. In the situation of two exposure categories (1 = unexposed, 2 = exposed), we define $\mathrm{RR}_1$ and $\mathrm{RR}_2$ as the relative risks in the unexposed and exposed strata, respectively. The relative risk ratio $\mathrm{RRR} = \mathrm{RR}_2/\mathrm{RR}_1$ is a measure of the extent of the risk associated with the allele, depending on the exposure status of the case. For instance, a ratio larger than 1 implies that an exposed child carrying the variant allele has a higher risk than the unexposed child carrying the variant allele.

The PoO effect can be seen as a statistical interaction between the transmitted allele and its parental origin, whereas the GxE effect is an interaction between a main fetal effect with an external environment. It is thus natural to consider a PoOxE effect as a two-way interaction that takes into account both parent of origin and environmental exposure in the same estimate. At a locus with two alleles and a dichotomous environmental exposure, the ratio

$$\mathrm{RRR} = (\mathrm{RR}_{M,2}/\mathrm{RR}_{F,2})/(\mathrm{RR}_{M,1}/\mathrm{RR}_{F,1}) \quad (2)$$

is the PoO effect in the second stratum compared with the PoO effect in the first stratum. If $\mathrm{RRR} = 1$, it means that there may well be PoO effects, but that they, when measured on a multiplicative scale, are the same in both environmental strata. Similarly, since Eqn (2) may also be expressed as

$$\mathrm{RRR} = (\mathrm{RR}_{M,2}/\mathrm{RR}_{M,1})/(\mathrm{RR}_{F,2}/\mathrm{RR}_{F,1}),$$

we will have $\mathrm{RRR} = 1$ if a GxE effect is the same for alleles of both parental origins. It is worth noting that the actual direction of an effect (i.e., $\mathrm{RRR} > 1$ or $\mathrm{RRR} < 1$) depends on which allele and exposure group are chosen as reference.

### 2.2.1 | The Wald test for interaction

In the log-linear model, statistical inference is performed on log-transformed relative risks and relative risk ratios. Thus, in the PoOxE situation, we would like to test the full interaction hypothesis

$$\beta_{M,1} - \beta_{F,1} = \beta_{M,2} - \beta_{F,2} = \cdots = \beta_{M,S} - \beta_{F,S},$$

where $\beta_{M,s}$ and $\beta_{F,s}$ are the log relative risks within stratum $s$, depending on whether the allele is derived from the mother or the father. Within each mutually exclusive exposure stratum, $s = 1, 2, \ldots, S$, we calculate $\hat{\beta}_s = \hat{\beta}_{M,s} - \hat{\beta}_{F,s}$, the difference between parental relative risks estimated on a log-scale. From the asymptotic theory of log-linear models (Christensen, 1997, Ch. 1 2.3), $\hat{\boldsymbol{\beta}}$ follows approximately a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance–covariance matrix $\boldsymbol{\Sigma}$,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_S \end{bmatrix} \sim \text{MVN}(\boldsymbol{\beta}, \boldsymbol{\Sigma}).$$

Because the strata are independent, the estimate of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_S^2 \end{bmatrix} = \text{diag}\left(\left[\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_S^2\right]\right),$$

where $\hat{\sigma}_s^2 = \hat{\sigma}_{M,s}^2 + \hat{\sigma}_{F,s}^2 - 2\hat{\rho}_{M,F,s}\hat{\sigma}_{M,s}\hat{\sigma}_{F,s}$, with $\hat{\rho}_{M,F,s}$ being the correlation between $\hat{\beta}_{M,s}$ and $\hat{\beta}_{F,s}$ within stratum $s$.

The Wald test can then be used to conduct post-hoc inference on the β parameters, based on the asymptotic normality (Agresti, 2013, Ch. 1.3). Let $\boldsymbol{D}$ be an appropriate $r \times S$ contrast matrix for the β parameters, with $r \leq S - 1$. It follows that asymptotically,

$$\boldsymbol{D}\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{D}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{D}}),$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{D}} = \boldsymbol{D}\hat{\boldsymbol{\Sigma}}\boldsymbol{D}^T$. The Wald test statistic is then

$$T = (\boldsymbol{D}\hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{D}}^{-1} (\boldsymbol{D}\hat{\boldsymbol{\beta}}).$$

Under the null hypothesis of $\boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{0}$, $T$ has an approximate chi-squared distribution with $r$ degrees of freedom, $\chi^2(r)$.

In the PoOxE test, our null hypothesis can be seen as a test of all strata $s = 2, \ldots, S$ against the first stratum $s = 1$; that is, the test takes the form

$$\boldsymbol{D}\boldsymbol{\beta} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \times \begin{bmatrix} \beta_{M,1} - \beta_{F,1} \\ \beta_{M,2} - \beta_{F,2} \\ \vdots \\ \beta_{M,S} - \beta_{F,S} \end{bmatrix} = 0.$$

Hence, the Wald test statistic has an approximate $\chi^2$ distribution with $r = S - 1$ degrees of freedom under the null hypothesis of no PoOxE effect. This is an overall test for any difference in PoO effects across strata when measured on a log risk scale.

Interactions with a continuous exposure variable can be incorporated in our framework by categorizing the variable into an appropriate number of categories and testing for a trend-type association of the resulting ordinal variable. This approach is outlined for GxE effects in Skare et al. (2012), and a test for trend is included in Haplin.

### 2.2.2 | PoOxE analysis of X-linked markers

Genetic association analyses of X-linked markers are especially relevant if the prevalence of a complex trait differs systematically for males and females. Various penetrance models in Haplin address different causal scenarios that apply to an X-linked disease locus. The models depend on the assumptions made regarding allele-effects in males versus females, and might include sex-specific baseline risks, shared or distinct relative risks for males and females, and X-inactivation in females. A detailed description of parameterization models is provided in a previous study (Jugessur et al., 2012b). Haplin also allows for PoOxE analyses of X-linked markers. Separate PoOxE analyses on males only are not possible; females are needed to obtain a contrast between maternally and paternally derived X-chromosome alleles. However, fathers and male children contribute to estimating allele frequencies, and importantly, to facilitate haplotype reconstruction. Relevant Haplin commands for analyzing PoOxE effects on the X-chromosome are provided in $S1$.

## 2.3 | Case triad study: Cleft palate–only data analysis

Cleft palate only (CPO) is a common craniofacial birth defect in humans, occurring with (nonisolated) or without (isolated) other congenital anomalies or identifiable malformation syndromes. The prevalence rate for isolated CPO is 5 per 10,000 births worldwide (Mossey & Castilla, 2003). A wide array of genetic variants and environmental risk factors have been reported to increase the risk of CPO (Mossey, Little, Munger, Dixon, & Shaw, 2009; Dixon, Marazita, Beaty, & Murray, 2011; Rahimov, Jugessur, & Murray, 2012). However, as with many other complex traits, the genetic variants discovered so far only explain a minor fraction of the phenotypic variability. From our previously published GWAS (Beaty et al., 2010, 2011; Shi et al., 2012), the genotypes for 1575 individuals from 550 isolated CPO families were available, including 466 complete case–parent triads. These families were mainly of European and Asian ancestry, but a small number of families of other ethnicities were also present.

We considered three SNPs from the GWAS data to illustrate our PoOxE approach. On these SNPs, we conducted pooled analyses using all ethnicities, as well as separate analyses for Europeans only. The environmental factor was maternal cigarette smoking during the periconceptional period, that is, from 3 months before conception until 3 months into pregnancy, a window of exposure of 6 months in total. In the self-administered questionnaire of the Norway Facial Clefts Study (https://www.niehs.nih.gov/research/atniehs/labs/epi/studies/ncl/index.cfm), this was evaluated as a simple yes/no response to ever having smoked during this period. The GWAS data set is available at the dbGAP database (http://www.ncbi.nlm.nih.gov/gap) under accession ID phs000094.v1.p1. Information on quality control and detailed characterizations of study participants and environmental exposure have been provided elsewhere (Haaland et al., 2017). Ethics approvals were obtained from the respective ethics committees for all the data in the cleft consortium. Background information on the study is provided in the original publication (Beaty et al., 2010).

## 3 | RESULTS

### 3.1 | Case triad study: Illustration of PoOxE data analysis

To illustrate our PoOxE test, we considered three SNPs from our GWAS data on CPO (Beaty et al., 2010, 2011; Shi et al., 2012). We only used top hits from previous studies, employing the same genetic triad data. Hence, the examples serve only as an illustration of our PoOxE test and not as independent replications of previous findings. Because our PoOxE approach integrates the PoO and GxE models, we start with examples of PoO effects (Table 1a) and GxE effects (Table 1b) before looking at the combined PoOxE effects (Table 1c).

The SNP rs7516430, located in the gene for "chromodomain helicase DNA binding protein 1-like" or *CHD1L* on chromosome 1, had one of the most distinct signals in a previous PoO GWAS analysis of CPO by Shi et al. (2012). We re-analyzed the data for this SNP on Europeans only, applying a Wald test. Table 1a (first row) presents the PoO estimates $\mathrm{RR}_M$, $\mathrm{RR}_F$ and $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F$. The most frequent allele, *A*, was used as reference. If allele *T* is inherited from the mother, it increases the risk of CPO. If, on the other hand, *T* is inherited from the father, the risk of CPO is nearly halved. As a result, RRR = 3.42. There is a qualitative PoO effect with *P*-value $5.6 \times 10^{-5}$. Note that the PoO effects were estimated without stratifying on the exposure, smoking. Hence, by assumption, the estimates do not differ between strata. We still included the corresponding rows in the table to facilitate comparison with the following analyses. Table 1a also includes tests for GxE and PoOxE effects for this SNP (second and third row, respectively). However, no significant interactions were found.

The SNP rs470563 is associated with a higher risk of CPO in the presence of maternal smoking (Beaty et al., 2011). It is located in the gene "zinc finger protein 236" (*ZNF*236) on chromosome 18, and the re-analyzed GxE results are presented in Table 1b (second row). Relative to allele *C*, allele *G* is associated with a decreased risk of CPO among smokers and an increased risk among nonsmokers. Consequently, RRR = 0.42, and this qualitative effect has a *P*-value of $4.5^{-4}$. It is important to note that although maternal smoking appears to be beneficial at first sight, this apparent risk-reducing effect of smoking is contingent on the choice of reference allele. Switching the reference and variant allele inverts the estimated value of the RRR. Obviously, the main effect of smoking cannot be assessed from case-triad designs alone, without independent controls. Therefore, the GxE RRR measures only how smoking *modifies* the estimated fetal genetic effects. For rs470563, we did not detect any significant PoO or PoOxE effects (Table 1b, first and third row, respectively). Note that the GxE effects were estimated without stratifying on parental origin. The columns in Table 1b, related to $\mathrm{RR}_M$ and $\mathrm{RR}_F$, are therefore equal by assumption.

In a separate study, we used the PoOxE test presented herein to perform a GWAS analysis of PoO interactions with maternal smoking and other exposures in Haplin (Haaland et al., 2017). The SNP rs2964137, located in the gene "interactor of little elongation complex ELL subunit 1" *(ICE1)*, had one of the strongest signals in our search for PoOxE effects, and the PoO, GxE, and PoOxE results are shown in Table 1c. The risk estimates are relative to allele G, which is the most frequent. For this SNP, there is no evidence of a PoO effect independent of strata (first row) or of any GxE effect for fetal genes independent of parental origin (second row). Nevertheless, we found a qualitative PoOxE effect, RRR = 0.09, with *P*-value $6.5 \times 10^{-7}$ (Table 1c, third row). The relative risk associated with allele C is nearly halved if derived from exposed mothers, and it is more than doubled if derived from exposed fathers. An opposite effect is seen in nonsmokers.

Haplin uses parallel processing of its analyses, and the run time of a GWAS analysis is therefore manageable. Our genome wide search for PoOxE effects was performed on Europeans only, comprising 762 individuals from 269 case families (mostly triads). Altogether 424,401 SNPs passed the quality controls and were included in our PoOxE analysis. We used eight CPU cores with 2.5 GHz per core, and the approximate run time of Haplin was 58 hours.

### 3.2 | Operating characteristics and small sample behavior of the PoOxE test

We investigated the performance of our PoOxE test by evaluating its power in various settings. Power and sample size can be computed from the asymptotic variance–covariance structure underlying the Wald test; this approach is implemented in

**TABLE 1** PoO, GxE and PoOxE effects for cleft palate-only example SNPs

| **a) rs7516430, *CHD1L*[1]** | | | | |
|---|---|---|---|---|
| Test effect | Stratum | $RR_M$ | $RR_F$ | $RR_M/RR_F$ |
| PoO effects* | $RR_S$ | 1.79 | 0.52 | 3.42 (1.86, 6.15) |
| | $RR_{NS}$ | 1.79 | 0.52 | 3.42 (1.86, 6.15) |
| | $RR_S/RR_{NS}$ | 1 (–) | 1 (–) | 1 (–) |
| GxE effects** | $RR_S$ | 1.22 | 1.22 | 1 (–) |
| | $RR_{NS}$ | 1.06 | 1.06 | 1 (–) |
| | $RR_S/RR_{NS}$ | 1.15 (0.51, 2.61) | 1.15 (0.51, 2.61) | 1 (–) |
| PoOxE effects | $RR_S$ | 1.88 | 0.66 | 2.83 (0.90, 8.63) |
| | $RR_{NS}$ | 1.76 | 0.48 | 3.68 (1.80, 7.37) |
| | $RR_S/RR_{NS}$ | 1.07 (0.43, 2.69) | 1.40 (0.40, 4.83) | 0.77 (0.20, 2.91) |
| **b) rs470563, *ZNF236*[2]** | | | | |
| Test effect | Stratum | $RR_M$ | $RR_F$ | $RR_M/RR_F$ |
| PoO effects* | $RR_S$ | 0.95 | 1.07 | 0.89 (0.67, 1.17) |
| | $RR_{NS}$ | 0.95 | 1.07 | 0.89 (0.67, 1.17) |
| | $RR_S/RR_{NS}$ | 1 (–) | 1 (–) | 1 (–) |
| GxE effects** | $RR_S$ | 0.48 | 0.48 | 1 (–) |
| | $RR_{NS}$ | 1.15 | 1.15 | 1 (–) |
| | $RR_S/RR_{NS}$ | 0.42 (0.26, 0.68) | 0.42 (0.26, 0.68) | 1 (–) |
| PoOxE effects | $RR_S$ | 0.44 | 0.52 | 0.86 (0.39, 1.87) |
| | $RR_{NS}$ | 1.09 | 1.22 | 0.89 (0.66, 1.20) |
| | $RR_S/RR_{NS}$ | 0.41 (0.21, 0.79) | 0.42 (0.23, 0.80) | 0.96 (0.41, 2.24) |
| **c) rs2964137, *ICE1*[3]** | | | | |
| Test effect | Stratum | $RR_M$ | $RR_F$ | $RR_M/RR_F$ |
| PoO effects* | $RR_S$ | 1.42 | 1.06 | 1.34 (0.90, 1.97) |
| | $RR_{NS}$ | 1.42 | 1.06 | 1.34 (0.90, 1.97) |
| | $RR_S/RR_{NS}$ | 1 (–) | 1 (–) | 1 (–) |
| GxE effects** | $RR_S$ | 1.16 | 1.16 | 1 (–) |
| | $RR_{NS}$ | 1.25 | 1.25 | 1 (–) |
| | $RR_S/RR_{NS}$ | 0.93 (0.54, 1.60) | 0.93 (0.54, 1.60) | 1 (–) |
| PoOxE effects | $RR_S$ | 0.53 | 2.57 | 0.21 (0.09, 0.46) |
| | $RR_{NS}$ | 1.88 | 0.85 | 2.22 (1.41, 3.43) |
| | $RR_S/RR_{NS}$ | 0.28 (0.13, 0.58) | 3.03 (1.45, 6.35) | 0.09 (0.04, 0.24) |

*PoO effects were estimated without stratifying on exposure. The rows corresponding to environmental strata are therefore equal by assumption.

**GxE effects were estimated without stratifying on parental origin. The columns related to $RR_M$ and $RR_F$ are therefore equal by assumption.

- The estimates are relative to the most frequent allele
- $RR_M$ and $RR_F$ are the relative risks depending on parental origin
- $RR_{NS}$ and $RR_S$ are the relative risks depending on exposure status (nonsmokers or smokers)

[1] Overall allele frequencies: A 0.88; T 0.12; Europeans only

[2] Overall allele frequencies: C 0.57; G 0.43; Whole sample

[3] Overall allele frequencies: G 0.52; C 0.48; Europeans only

Haplin. The Haplin framework also includes a complete setup for power calculations through simulations, which is a robust way of checking software implementations, power, small-sample behavior, and attained significance level. A detailed derivation of our asymptotic approximation formulae is given in Appendix A.2. Relevant example code for power calculations in Haplin is provided in S1.

We examined the power of the PoOxE test using the above-mentioned asymptotic approximations. We first analyzed the power for a single SNP at the 5% nominal significance level. Power calculations for increasing relative risk ratios, RRRs, are shown in Figure 1. For simplicity, we set $RR_{M,1} = RR_{F,1} = RR_{F,2} = 1$ in all scenarios so that the value of RRR in Equation (2) is equal to the value of $RR_{M,2}$. Moreover, we assumed equally sized exposed and unexposed groups. The left panel of Figure 1 shows the statistical power for an increasing number of case–parent triads and a minor allele frequency (MAF) of 0.2. The black solid line is equal in all panels and is based on

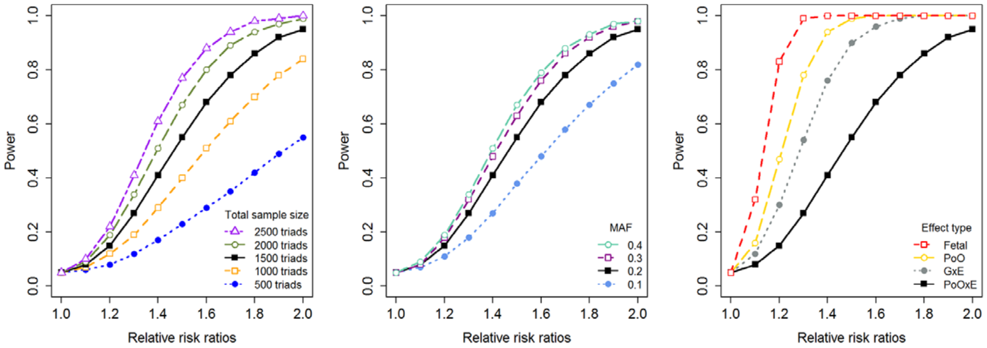

**FIGURE 1** Single-SNP power analysis for the PoOxE test for increasing relative risk ratios (increasing values of $RR_{M,2}$; $RR_{M,1} = RR_{F,1} = RR_{F,2} = 1$) at the 0.05 nominal significance level. Equally sized exposure groups are assumed. Left panel: Increasing number of case–parent triads, and MAF = 0.2; Middle panel: Increasing MAFs, and a total of 1500 case–parent triads; Right panel: Power comparison of the PoOxE, GxE (increasing values of $RR_2$; $RR_1 = 1$), PoO (increasing values of $RR_M$; $RR_F = 1$), and fetal effect (increasing values of RR) tests, MAF = 0.2, and a total of 1500 case–parent triads [Colour figure can be viewed at wileyonlinelibrary.com]

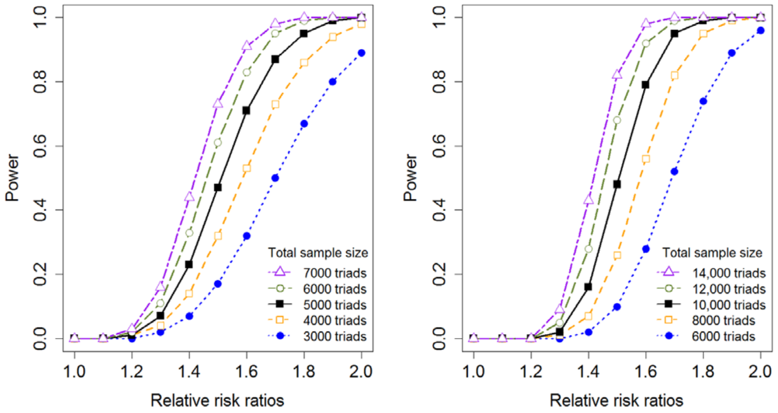

**FIGURE 2** GWAS power analysis for the PoOxE test for increasing relative risk ratios (increasing values of $RR_{M,2}$; $RR_{M,1} = RR_{F,1} = RR_{F,2} = 1$) and increasing number of case-parent triads, assuming equally sized exposure groups and MAF = 0.2. Left panel: Nominal significance level $10^{-4}$; right panel: Nominal significance level $5 \times 10^{-8}$ [Colour figure can be viewed at wileyonlinelibrary.com]

a total of 1500 case–parent triads, that is, 750 case–parent triads in both exposure categories. The middle panel depicts the power for increasing MAFs, using a total of 1500 case–parent triads. The right panel compares the power for various disease mechanisms (PoOxE, GxE, PoO, and fetal effects), using a total of 1500 case–parent triads and MAF = 0.2. Here, the fetal genetic effect is the direct risk associated with the child's allele, regardless of parent of origin or environmental exposures.

The power to detect PoOxE effects for a single SNP is sufficient for RRRs above 1.6–1.7 and a total sample size of 1500 case–parent triads with equally sized exposure groups. Nevertheless, larger sample sizes are needed if the MAF < 0.2 or if the ratio of exposed versus unexposed is highly skewed (the latter result is not shown). Because the PoOxE test stratifies on both parent of origin and exposure, detecting a PoOxE effect requires a larger sample size than detecting a PoO effect or a GxE effect. Naturally, greatest power is achieved in a search for fetal effects.

We also examined the power using nominal significance levels more relevant to GWAS settings. Figure 2 shows power analyses for increasing RRRs (i.e., increasing values of $RR_{M,2}$) with nominal significance levels $10^{-4}$ (left panel) and $5 \times 10^{-8}$ (right panel). The power is demonstrated for an increasing number of case–parent triads using equally sized exposure groups and a MAF of 0.2. With a nominal significance level of $10^{-4}$, approximately 5000 case–parent triads are required to detect RRRs of 1.6–1.7 with 80% power.

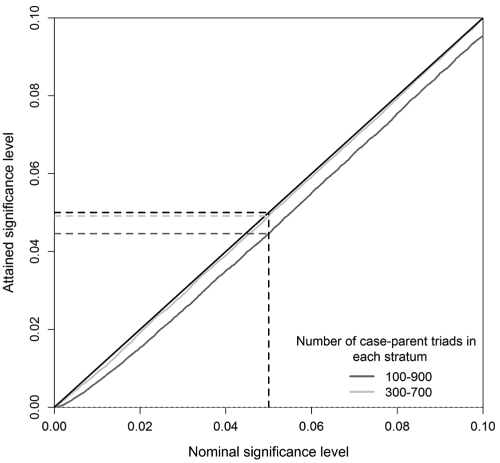

**FIGURE 3** Simulated $P$-values under the null hypothesis of no PoOxE effects based on 100,000 replications of data sets. The cumulative density plots compare the attained significance level with an expected uniform distribution under the null hypothesis (diagonal sloping line). A total of 1000 case–parent triads were divided into two exposure strata, and a MAF of 0.2 was assigned throughout. The distribution of case-parent triads in each stratum was as follows: 100–900 (dark grey line) and 300–700 (light grey line). If no bias is present, the observed significance levels should equal the nominal level of 0.05 (black dashed lines). The dark and light grey dashed horizontal lines show the attained significance levels corresponding to the simulated scenarios

With a nominal significance level of $5 \times 10^{-8}$, a sample size of 10,000 case-parent triads suffices for RRRs above 1.6.

Our PoOxE test is asymptotically unbiased. However, the asymptotic approximations underlying log-linear models may be suboptimal when the number of cases or controls is too small in one or more strata. When testing for GxE and PoOxE effects, one may occasionally encounter highly skewed exposure distributions. For example, in our CPO example, only 8 women of Asian ancestry answered “yes” to the question of maternal smoking during pregnancy, whereas the remaining 245 answered “no.” In such situations, the nominal significance level of the tests may be incorrect; the actual significance level is most easily assessed through simulations.

In Figure 3, cumulative density plots were used to examine the attained significance level of our PoOxE test. We obtained $P$-values from 100,000 simulated data sets under the null hypothesis ($\mathrm{RR}_{M,1} = \mathrm{RR}_{M,2} = \mathrm{RR}_{F,1} = \mathrm{RR}_{F,2} = 1$). The $P$-values should be uniformly distributed when the null hypothesis is true. Hence, if no bias is present, the $P$-values would fall close to the diagonal line. Throughout, a total of 1000 case–parent triads were divided into two exposure groups, and an MAF of 0.2 was assigned to both strata. Two scenarios were investigated according to the distribution of exposed and unexposed triads. In the first scenario (100–900), the smallest stratum comprised 100 case–parent triads. In the second scenario (300–700), the smallest stratum comprised 300 case–parent triads.

As expected, we observed a small bias for the PoOxE test when the number of cases in one exposure group was low, obtaining larger $P$-values than expected. At the 0.05 nominal level, the attained significance level is 0.045 in the 100–900 setting. For lower significance levels, typically occurring in genome wide analyses, this bias might become substantial. Each exposure group should be large enough so that the asymptotic approximation of the estimator, $\hat{\boldsymbol{\beta}}$, is sufficiently precise. Hence, the bias would be less pronounced for skewed exposure distributions at larger sample sizes (such as in a 1000–9000 setting). In other words, the unbalanced exposure design itself is not the cause of the observed deflation. The bias is negligible in the 300–700 setting, verifying that our PoOxE test attains the nominal significance level when the sample size of the smallest stratum increases.

## 4 | CONCLUDING REMARKS

In this study, we have proposed a statistical method for detecting PoOxE effects. Postestimation in the log-linear framework, incorporated into the Haplin software, allows us to combine the theory on PoO and GxE effects to test for the second-order PoOxE effect. Although PoO and GxE studies abound, the combination has hardly been analyzed, in spite of its obvious biological relevance. Wang et al. (2011) proposed an interesting test to screen for interactions between imprinted genes and environmental exposures in a more restricted setting than our approach. Specifically, when testing for imprinted genes, Wang et al. assume that either the maternally or the paternally inherited allele is silenced so that only the other allele has an effect. This is in contrast to our PoO effect, which measures the difference between the effects of maternally and paternally derived alleles. Although the assumption of imprinted genes may increase testing power when it is true, it has the drawback of being more easily confused with ordinary fetal effects. For instance, if $\mathrm{RR}_M = \mathrm{RR}_F = 1.5 > 1$, this would trigger a test for imprinted genes but not for PoO.

Wang et al. (2011) use conditional logistic regression to analyze birth cohort designs with mother–offspring pairs. Our log-linear framework is a general approach to the full hybrid design with complete or incomplete case triads possibly combined with control triads. We are therefore able to separate the effects of maternal alleles from the effect of maternally derived fetal alleles, which is particularly important in perinatal epidemiology, where the phenotype of the fetus can be influenced by either of the two sources (Hager, Cheverud, & Wolf, 2008). Additionally, our model provides a full maximum likelihood setup that allows us to estimate allele frequencies, haplotyping of multiple SNPs, and imputation of missing genotypes. Ambiguous (heterozygous) mother–offspring combinations need not be excluded as in the conditional logistic setup; they incorporate naturally into the model and provide data for the allele frequency estimation. Similarly, within the Haplin framework, PoOxE effects may also be detected on the X-chromosome, where female offspring provide a contrast between maternally and paternally derived alleles; fathers and male offspring contribute to allele frequency estimation and precise haplotyping (Jugessur et al., 2012b). Finally, the data handling in Haplin enables a full genome-wide screen for PoOxE effects.

Detailed study planning typically requires calculating the sample sizes needed to obtain sufficient power. Because statistical power depends on multiple factors including haplotype frequencies, penetrance model, and so on, published power tables for genetic studies are typically too restrictive, and software often covers only basic genetic models. As illustrated in S1, Haplin provides extensive power simulations, even covering the complex setup of PoOxE analyses. By entering the necessary parameters, the user can easily perform either "raw" simulations of power or use a very fast power calculation based on the asymptotic distribution of the parameter estimates.

In a GWAS analysis, the power to detect PoOxE effects is generally low. However, a candidate gene approach would reduce the complexity of multiple comparisons and enable a search for PoOxE effects when the sample size is limited. Specific environmental exposures that relate directly to the putative cause of the PoO effect of a candidate gene should be used in a PoOxE test. For example, one might assume that a detected PoOxE effect has a better chance of revealing a causal relationship involving genomic imprinting due to methylation than the standard PoO or GxE searches. A selection of relevant candidate genes might therefore be based on a GWAS screen for PoO or GxE effects.

Tracking the different etiologic mechanisms underlying complex diseases is crucial in improving diagnosis, prognosis, and prevention. The test for PoOxE effects and the comprehensive framework for assessing statistical power for genetic association analyses presented in this article are thus important contributions in advancing our understanding of the different etiologic mechanisms that underlie complex traits.

## 5 | ELECTRONIC DATABASE INFORMATION

Haplin is implemented as a standard package in the statistical software ***R*** (R Core Team, 2016) and can be installed from the official R package archive, CRAN (https://cran.r-project.org). Our website (http://folk.uib.no/gjessing/genetics/software/haplin) provides further information.

## ACKNOWLEDGEMENTS

The authors thank Prof. Ivar Heuch for his valuable comments.

## AUTHORS' CONTRIBUTIONS

Contribution of analytic tools and method development: M. G., J. R., H. K. G.; Data analysis: M. G., Ø. A. H., R. T. L., A. J., H. K. G.; Manuscript preparation: M. G., Ø. A. H., J. R., R. T. L., A. J., H. K. G.

## CONFLICT OF INTEREST

The authors declare that they have no competing interests.

## ORCID

*Øystein A. Haaland* http://orcid.org/0000-0001-5288-7879
*Astanand Jugessur* http://orcid.org/0000-0002-2604-2132
*Håkon K. Gjessing* http://orcid.org/0000-0002-3544-1063

## REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Bartolomei, M. S., & Tilghman, S. M. (1997). Genomic imprinting in mammals. *Annual Review of Genetics*, *31*, 493–525.

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., ... Scott, A. F. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genetics*, *2*, 525–529.

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

Christensen, R. (1997). *Log-linear models and logistic regression* (2nd ed.). New York: Springer.

Connolly, S., & Heron, E. A. (2014). Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Briefings in Bioinformatics*, *16*, 429–448.

Cordell, H. J. (2004). Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genetic Epidemiology*, *26*, 186–205.

Cordell, H. J., Barratt, B. J., & Clayton, D. G. (2004). Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genetic Epidemiology*, *26*, 167–185.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *39*, 1–38.

Dixon, M. J., Marazita, M. L., Beaty, T. H., & Murray, J. C. (2011). Cleft lip and palate: Understanding genetic and environmental influences. *Nature Reviews Genetics*, *12*, 167–178.

Gjessing, H. K., & Lie, R. T. (2006). Case-parent triads: Estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Annals of Human Genetics*, *70*, 382–396.

Guilmatre, A., & Sharp, A. J. (2012). Parent of origin effects. *Clinical Genetics*, *81*, 201–209.

Haaland, Ø. A., Jugessur, A., Gjerdevik, M., Romanowska, J., Shi, M., Beaty, T. H., ... Gjessing, H. K. (2017). Genome-wide analysis of parent-of-origin interaction effects with environmental exposure (POOxE): An application to European and Asian cleft palate trios. *PLoS One*, *12*, e0184358.

Hager, R., Cheverud, J. M., & Wolf, J. B. (2008). Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics*, *178*, 1755–1762.

Howey, R., Mamasoula, C., Töpf, A., Nudel, R., Goodship, J. A., Keavney, B. D., & Cordell, H. J. (2015). Increased power for detection of parent-of-origin effects via the use of haplotype estimation. *American Journal of Human Genetics*, *97*, 419–434.

Jugessur, A., Skare, Ø., Harris, J. R., Lie, R. T., & Gjessing, H. K. (2012a). Using offspring-parent triads to study complex traits: A tutorial based on orofacial clefts. *Norsk Epidemiologi*, *21*, 251–267.

Jugessur, A., Skare, Ø., Lie, R. T., Wilcox, A. J., Christensen, K., Christiansen, L., ... Gjessing, H. K. (2012b). X-linked genes and risk of orofacial clefts: Evidence from two population-based studies in Scandinavia. *PLoS One*, *7*, 1–12.

Knapp, M., Seuchter, S. A., & Baur, M. P. (1993). The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *American Journal of Human Genetics*, *52*, 1085–1093.

Lawson, H. A., Cheverud, J. M., & Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. *Nature Reviews Genetics*, *14*, 609–617.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–53.

Mossey, P. A., & Castilla, E. E. (2003). *Global registry and database on craniofacial anomalies*. Geneva: World Health Organization.

Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J., & Shaw, W. C. (2009). Cleft lip and palate. *Lancet*, *374*, 1773–1785.

Pasaniuc, B., & Price, A. L. (2016). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics*, *18*, 117–127.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rahimov, F., Jugessur, A., & Murray, J. C. (2012). Genetics of nonsyndromic orofacial clefts. *Cleft Palate-Craniofacial Journal*, *49*, 73–91.

Reik, W., & Walter, J. (2001). Genomic imprinting: Parental influence on the genome. *Nature Reviews Genetics*, *2*, 21–32.

Schaid, D. J., & Sommer, S. S. (1993). Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics*, *53*, 1114–1126.

Shi, M., Christensen, K., Weinberg, C. R., Romitti, P., Bathum, L., Lozada, A., ... Murray, J. C. (2007). Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *American Journal of Human Genetics*, *80*, 76–90.

Shi, M., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., ... Beaty, T. H. (2012). Genome wide study of maternal and parent-of-origin effects on the etiology of orofacial clefts. *American Journal of Medical Genetics Part A*, *158 A*, 784–794.

Shi, M., Umbach, D. M., & Weinberg, C. R. (2010). Testing haplotype-environment interactions using case-parent triads. *Human Heredity*, *70*, 23–33.

Skare, Ø., Jugessur, A., Lie, R. T., Wilcox, A. J., Murray, J. C., Lunde, A., ... Gjessing, H. K. (2012). Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. *Annals of Human Genetics*, *76*, 221–236.

Wang, S., Yu, Z., Miller, R. L., Tang, D., & Perera, F. P. (2011). Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Human Heredity 71*, 196–208.

Weinberg, C. R., & Umbach, D. M. (2005). A hybrid design for studying genetic influences on risk of diseases with onset early in life. *American Journal of Human Genetics*, *77*, 627–636.

Weinberg, C. R., Wilcox, A. J., & Lie, R. T. (1998). A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *American Journal Human Genetics*, *62*, 969–978.

Wilcox, A. J., Weinberg, C. R., & Lie, R. T. (1998). Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." *American Journal of Epidemiology*, *148*, 893–901.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Gjerdevik M, Haaland ØA, Romanowska J, et al. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. *Ann Hum Genet*. 2018;82:60–73. https://doi.org/10.1111/ahg.12224

## APPENDIX A

### A.1 PoOxE effects in the haplotype situation

The majority of existing methods to investigate PoO and GxE effects are performed using a single-marker approach in which each SNP is analyzed individually. However, haplotype analysis should enhance the possibility of "bracketing" a causal variant if the haplotype has a SNP on each side of the variant. The theory of PoOxE effects for the single-marker setting can easily be extended to haplotypes. We here present a detailed derivation of the PoOxE test.

We assume a multiplicative dose–response effect and a reference haplotype approach. Without loss of generality, the first haplotype in arbitrary order is chosen as reference. Let $H$ denote the number of haplotypes and $S$ the number of independent exposure strata. We define $\hat{\boldsymbol{\beta}}_{M,s} = [\hat{\beta}_{2,M,s}, \hat{\beta}_{3,M,s}, \dots, \hat{\beta}_{H,M,s}]^T$ and $\hat{\boldsymbol{\beta}}_{F,s} = [\hat{\beta}_{2,F,s}, \hat{\beta}_{3,F,s}, \dots, \hat{\beta}_{H,F,s}]^T$, the relative risk estimates on a log-scale for each haplotype within exposure stratum $s$ ($s = 1, 2, \dots, S$), depending on parental origin. We calculate the difference $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_{M,s} - \hat{\boldsymbol{\beta}}_{F,s}$ and the corresponding asymptotic variance–covariance estimate

$$\hat{\boldsymbol{\Sigma}}_s = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{M,s} & \hat{\boldsymbol{\Sigma}}_{M,F,s} \\ \hat{\boldsymbol{\Sigma}}_{M,F,s} & \hat{\boldsymbol{\Sigma}}_{F,s} \end{bmatrix},$$

in which each element is a combined $(H-1) \times (H-1)$ variance–covariance matrix for haplotypes $2, 3, ..., H$.

We would like to test the null hypothesis

$$\boldsymbol{\beta}_{M,1} - \boldsymbol{\beta}_{F,1} = \boldsymbol{\beta}_{M,2} - \boldsymbol{\beta}_{F,2} = \cdots = \boldsymbol{\beta}_{M,S} - \boldsymbol{\beta}_{F,S}.$$

This can be reformulated as

$$\boldsymbol{D\beta} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{0} & -\boldsymbol{I} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots & -\boldsymbol{I} \end{bmatrix} \times \begin{bmatrix} \boldsymbol{\beta}_{M,1} - \boldsymbol{\beta}_{F,1} \\ \boldsymbol{\beta}_{M,2} - \boldsymbol{\beta}_{F,2} \\ \vdots \\ \boldsymbol{\beta}_{M,S} - \boldsymbol{\beta}_{F,S} \end{bmatrix} = \boldsymbol{0}.$$

Here, $\boldsymbol{I}$ is the $(H-1) \times (H-1)$ identity matrix. From basic asymptotic theory of log-linear models, we have that asymptotically

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_S \end{bmatrix} \sim MVN(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where

$$\hat{\boldsymbol{\Sigma}} = \text{diag}\left(\left[\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2, \dots, \hat{\boldsymbol{\Sigma}}_S\right]\right).$$

Consequently, under the null hypothesis, the Wald statistic, $T = (\boldsymbol{D}\hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{D}}^{-1} (\boldsymbol{D}\hat{\boldsymbol{\beta}})$, has an approximate $\chi^2$ distribution with $(H-1)(S-1)$ degrees of freedom.

### A.1.1 Haplotype example

Our Haplin framework allows a straightforward PoOxE analysis of haplotypes. As an illustration, we formed haplotypes by using one SNP on each side of the previously analyzed SNP rs2964137 in *ICE1* (i.e., rs2964447-rs2964137-rs6868526). We excluded haplotypes with frequencies below 1%, which left us with three haplotypes for our analysis. The results are displayed in Table 2, and the risk estimates are relative to the reference A-C-C haplotype. The first two SNPs are in strong linkage disequilibrium ($r^2 = 0.996$); the first SNP is therefore redundant and the same information can be obtained by using only the two last SNPs ($r^2 = 0.427$). Both the T-G-C and T-G-G haplotypes display PoOxE effects when analyzed separately against the reference, using the Wald test with one degree of freedom ($P$-value $= 2.1 \times 10^{-5}$ and $P$-value $= 9.9 \times 10^{-4}$). The PoOxE effect is stronger when both haplotypes are analyzed jointly, with 2 degrees of freedom ($P$-value $= 8.5 \times 10^{-6}$). The separate relative risk estimates are fairly similar for the two haplotypes, indicating that the haplotype risks are driven by rs2964447 and rs2964137, which have the largest individual effect.

The joint haplotype analysis loses some power compared to the single-SNP analysis of rs2964137 due to haplotype reconstruction ($P$-value $8.5 \times 10^{-6}$ versus $6.5 \cdot 10^{-7}$). Moreover, the Wald test statistic has 2 degrees of freedom. Nonetheless, we do not know a priori which approach, single-marker or haplotype, will have the best likelihood of identifying an association.

## A.2 Statistical power

The power of a genetic association analysis depends on numerous factors, such as significance level, allele/haplotype frequencies, effect size, and family design. A sample size calculation will typically involve computing the number of families needed to be genotyped to achieve a preset power for a given effect size. For instance, one might wish to achieve 80% power to detect a fetal effect of RR = 2. The standard simulation approach to power calculations is the following. First, a sufficiently large number of data sets is simulated with appropriate parameter choices, such as effect size, sample size, family design, and so on. Then, the test is performed on each data set, and the power is the proportion of rejected null hypotheses. For a range of disease mechanisms, including PoO, GxE, and PoOxE effects, such power simulations are readily done in Haplin through the functions `hapRun` and `hapPower`. Relevant example code is provided in S1.

"Brute-force" simulations are especially useful for small to moderate data sets. In such situations, only simulation studies can indicate the extent and direction of the possible bias. Nevertheless, both power and sample size can be computed much more efficiently directly from the asymptotic distributions underlying the Wald test. Such calculations have been implemented for a number of genetic effects in the Haplin function `hapPowerAsymp`. The principles behind the asymptotic calculations are standard; we will in the following paragraphs outline the specifics of our model implementations.

All tests described in this paper are performed as Wald tests, using the asymptotic normal distribution of the log-scale parameters. In general, the power $\gamma$ of the Wald test with level $\alpha$ is

$$\gamma = 1 - F_{r,\lambda}(\chi^2_\alpha(r)), \tag{A.1}$$

**TABLE 2** PoOxE effects for cleft palate–only example haplotypes

| **rs2964447-rs2964137-rs6868526, *ICE1*** | | | | |
|---|---|---|---|---|
| Haplotype | Stratum | $RR_M$ | $RR_F$ | $RR_M/RR_F$ |
| T-G-C | $RR_S$ | 1.99 | 0.49 | 4.04 (1.75, 9.25) |
| | $RR_{NS}$ | 0.52 | 1.04 | 0.50 (0.31, 0.82) |
| | $RR_S/RR_{NS}$ | 3.79 (1.74, 8.22) | 0.47 (0.21, 1.05) | 7.98 (3.07, 20.77) |
| T-G-G | $RR_S$ | 1.30 | 0.24 | 5.35 (1.51, 18.19) |
| | $RR_{NS}$ | 0.68 | 1.30 | 0.52 (0.29, 0.96) |
| | $RR_S/RR_{NS}$ | 1.89 (0.70, 5.07) | 0.19 (0.06, 0.62) | 10.13 (2.55, 40.19) |

-Reference haplotype: A-C-C

-Overall haplotype frequencies: A-C-C 0.48; T-G-C 0.36; T-G-G 0.16; Europeans only

-$RR_M$ and $RR_F$ are the relative risks depending on parental origin.

-$RR_{NS}$ and $RR_S$ are the relative risks depending on exposure status (nonsmokers or smokers)

where $\chi^2_\alpha(r)$ is the α quantile of the chi-squared distribution with $r$ degrees of freedom, $F_{r,\lambda}$ is the cumulative distribution function of a noncentral chi-squared distribution $\chi^2(r,\lambda)$, and λ is the noncentrality parameter. To compute λ, consider first the simplest situation where we estimate a single effect, such as a fetal gene effect or a parent-of-origin effect, within a single stratum. Let $n$ be the number of case children in the stratum. As $n$ changes, we assume the composition of family structures within the stratum remains the same, relatively speaking. That is, we assume the ratio of control families to case families, the ratio of case mother–child dyads to complete case triads and so on, all remain the same. As before, we assume $\beta = \log(\mathrm{RR})$ is the log effect size in the stratum, and $\sigma^{(n)}$ is the standard error of $\hat{\beta}$ when estimated from all data in the stratum, with $n$ case children. If the family structures are kept fixed as $n$ increases, observe that $\sigma^{(n)} \approx \omega/\sqrt{n}$, where ω is the asymptotic standard error computed from the Fisher information in the maximum likelihood model. The value of ω is scaled to correspond to a sample with only one case child ($n = 1$) in a stratum. For instance, in a setting with 200 case triad and 100 control triads, ω would, theoretically, correspond to a stratum with one case triad and half a control triad. Note that the ω parameter typically depends in a relatively complex way on the family design and allele/haplotype frequencies, and also on the effect sizes.

The noncentrality parameter λ is then the squared standardized log effect size (Agresti, 2013, Ch. 6.6), that is,

$$\lambda = \left( \frac{\log(\mathrm{RR})}{\omega/\sqrt{n}} \right)^2 . \tag{A.2}$$

When the value of ω, corresponding to the appropriate model, has been determined, the power γ for a given sample size $n$ is readily computed from Eqn (A.1), with $r = 1$ and using the λ value computed from Eqn (A.2). Equivalently, for a given power γ, the necessary sample size can be computed by first finding the corresponding non-centrality parameter λ from Eqn (A.1), and then solving Eqn (A.2) for $n$ to obtain

$$n = \lambda\omega^2 / \log^2(\mathrm{RR}). \tag{A.3}$$

The relationship between γ and λ is illustrated in Figure 4 when $r = 1$. Note that the lower significance levels are relevant in situations where multiple testing must be accounted for.

### A.2.1 Sample size calculation for the PoO test

To ease the derivation of sample size estimation for the PoOxE test, we first illustrate the approach for our PoO test. When searching for PoO effects in a diallelic situation, the test statistic has one degree of freedom. Equations (A.1), (A.2), and (A.3) apply, with $\mathrm{RR} = \mathrm{RR}_M/\mathrm{RR}_F$. To facilitate power calculations "by hand" in simple situations, Table S1 provides the values of ω for selected PoO settings. Without loss of generality, in the following examples and derivations, we let the first allele in arbitrary order be the reference, with allele frequency $1 - P$. Note that if $P > 0.5$, the reference allele is the minor allele.

Consider an example of sample size calculation for the PoO test. Let $\mathrm{RR}_M = 2$, $\mathrm{RR}_F = 1$, and $P = 0.1$. From Table S1, we find that $\omega^2 = 19.5$. With level $\alpha = 0.05$ and desired power $\gamma = 80\%$, Figure 4 yields $\lambda = 7.85$. Applying Eqn (A.3), we need roughly 320 case–parent triads or, equivalently, 344 case–mother dyads or 404 case–father dyads (the $\omega^2$ values for case–father dyads are not included in Table S1). Note that the values of $\omega^2$ depend not only on the ratio RR but also on the individual values of both $\mathrm{RR}_M$ and $\mathrm{RR}_F$. These calculations can be verified directly by power calculations in Haplin, as shown in S1.

Although a limited selection of values of $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are included in Table S1, several symmetry relationships allow us to use the simple approach also in other scenarios. The power for testing PoO effects in case–parent triads for $\mathrm{RR}_M = x$ and $\mathrm{RR}_F = y$ is the same as when $\mathrm{RR}_M = y$ and $\mathrm{RR}_F = x$. Moreover, the power for testing PoO effects in triads if

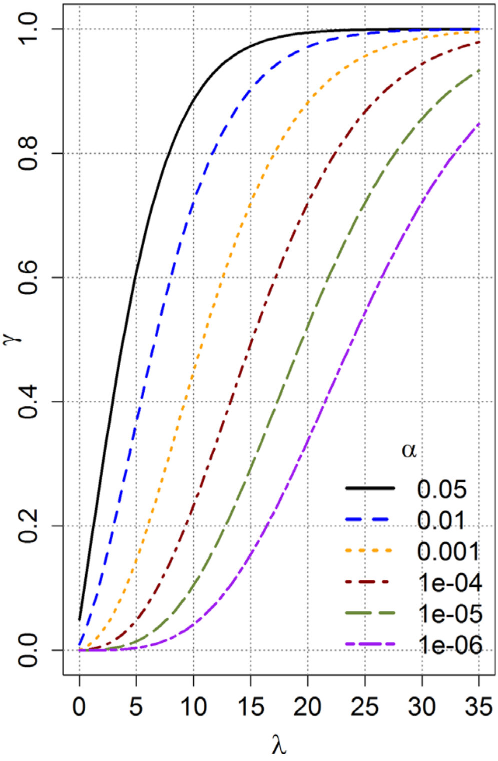

**FIGURE 4** Power, γ, as a function of the noncentrality parameter, λ, for differing values of the nominal significance level, α. Here, $\lambda = (\frac{\log(\mathrm{RR})}{\omega/\sqrt{n}})^2$, where log(RR) is the log effect size, $n$ is the number of case children, and ω is the asymptotic standard error of the log-parameter. The number of degrees of freedom is equal to 1 [Colour figure can be viewed at wileyonlinelibrary.com]

$\mathrm{RR}_M = x, \mathrm{RR}_F = y$, and $P = p$ is identical to the power when $\mathrm{RR}_M = 1/x$, $\mathrm{RR}_F = 1/y$, and $P = 1 - p$. Finally, testing for PoO effects in case–mother dyads for $\mathrm{RR}_M = x$, $\mathrm{RR}_F = y$, and $P = p$ is equivalent to testing for PoO effects in case–father dyads when $\mathrm{RR}_M = 1/y$, $\mathrm{RR}_F = 1/x$ and $P = 1 - p$.

### A.2.2 Sample size calculation for the PoOxE test

We now consider two independent strata with sample size (number of case children) $n_1$ and $n_2$, respectively, where we want to compare $\mathrm{RR}_1 = \mathrm{RR}_{M,1}/\mathrm{RR}_{F,1}$ in the first stratum with $\mathrm{RR}_2 = \mathrm{RR}_{M,2}/\mathrm{RR}_{F,2}$ in the second stratum. The variance of $\beta = (\beta_{M,2} - \beta_{F,2}) - (\beta_{M,1} - \beta_{F,1})$ is $\sigma_1^2 + \sigma_2^2$, where $\sigma_1^2 \approx \omega_1^2/n_1$ and $\sigma_2^2 \approx \omega_2^2/n_2$ are the variances in the first and second stratum, respectively. The power to detect PoOxE effects is thus fully determined by the power to assess PoO effects in each stratum. Given power γ, significance level α, the stratum-specific effects $\mathrm{RR}_1$ and $\mathrm{RR}_2$, and allele frequencies $P_1$ and $P_2$, as well as the ratio of sample sizes in the two strata, $\delta = n_2/n_1$, the PoOxE sample size calculation can be summarized in the following procedure:

1. Calculate $\omega_1^2$ and $\omega_2^2$ for the two exposure strata.
2. Calculate the sample size in the second stratum from the formula
$$n_2 = \frac{\lambda(\delta\omega_1^2 + \omega_2^2)}{\log^2(\mathrm{RR}_2/\mathrm{RR}_1)},$$
where λ corresponds to the power γ.
3. Calculate the sample size in the first stratum, $n_1 = n_2/\delta$.

Note that with two exposure strata, the number of degrees of freedom still equals one.

As an example, let $\mathrm{RR}_1 = 1$, $P_1 = 0.3$, $\mathrm{RR}_2 = 2.5$, and $P_2 = 0.1$, assuming $\mathrm{RR}_F = 1$ in both strata. For a given disease and environmental exposure, assume that it is reasonable to recruit twice as many case-parent triads in the first stratum as in the second (i.e., $\delta = 1/2$). From Table S1a, we find that $\omega_1^2$ =12.1 and $\omega_2^2$ =18.6. Hence, it is sufficient to enroll approximately 460 triads in the first stratum and 230 triads in the second stratum to achieve 80% power at the 5% nominal significance level. The full power calculations for PoOxE effects have also been implemented in the Haplin function `hapPowerAsymp`.

# S1. Haplin Commands

This section provides Haplin commands for analyzing PoO, GxE and PoOxE effects on candidate genes. We also show how a PoOxE analysis can be done on the X-chromosome. Relevant commands for power calculations are given, using both the asymptotic properties of the log-linear model and simulations. Haplin outputs are shown for selected examples. For a thorough description of the Haplin functions and their arguments, please refer our website at `http://folk.uib.no/gjessing/genetics/software/haplin`.

## PoO, GxE and PoOxE Analyses on Candidate Genes

The fictive example file “data.dat” contains data on three SNPs in the native Haplin data format, although other data formats, including standard pedigree files, can easily be used for our analyses. Each line in the file represents a case-parent triad, with missing data on parents coded as NA. Information on maternal smoking is included as a covariate in the first column, followed by columns containing the genetic data.

Our PoO examples are analyzed in Haplin using commands similar to

```
res.PoO <- haplin(filename = "data.dat",
                  markers = 2, n.vars = 1,
                  design = "triad", poo = T,
                  response = "mult", reference = "ref.cat",
                  use.missing = T)
```

We here analyze the second SNP in the data set (`markers = 2`), and there is only one column in the data file to the left of the genetic data (`n.vars = 1`). The standard case-parent triad design without independent controls is specified by `design = "triad"`. The argument `poo = T` enables estimation of PoO effects. A multiplicative dose-response model is specified by `response = "mult"`;

reference = "ref.cat" chooses the most frequent allele/haplotype as reference. When use.missing is set to true, Haplin uses the EM algorithm to obtain risk estimates, accounting for incomplete triads. Note that both PoO and maternal risks, controlling for possible confounding with one another, are estimated simultaneously by including maternal = T. The most relevant output is tabulated by the command haptable(res.PoO).

The GxE effects are calculated by a two-step procedure. First, the genetic effects in each stratum of the environmental exposure are estimated using the function haplinStrat:

```
res.GxE <- haplinStrat(filename = "data.dat",
                markers = 2, n.vars = 1,
                strata = 1, design = "triad",
                poo = F, response = "mult",
                reference = "ref.cat", use.missing = T)
```

The exposure covariate is indicated by the argument strata. Second, the results from all strata are compared with the Wald test using the function gxe(res.GxE). Important output for each stratum is obtained by haptable(res.GxE).

Our PoOxE effects are estimated by similar commands to the GxE analyses. However, the argument poo must be set to true in haplinStrat. Haplin then computes the PoO effects within each stratum before contrasting the results through the function gxe. We here show an example of PoOxE analysis in the haplotype situation. The haplotypes are readily specified in Haplin through the argument markers, which in our case is formed by the first, second and third SNP in the data.

```
res.PoOxE <- haplinStrat(filename = "data.dat",
                markers = c(1,2,3), n.vars = 1,
                strata = 1, design = "triad",
                poo = T, response = "mult",
                reference = "ref.cat", use.missing = T)


gxe(res.PoOxE)
    gxe.test      chisq df         pval
1 haplo.freq  0.6910991  2 7.078313e-01
2        poo 23.3532786  2 8.489849e-06
```

For each test that is performed, the output shows the Wald chi-squared test value, degrees-of-freedom and the resulting p-value. Results for the haplotype frequency is always displayed in the first row. We are interested in the PoOxE results, which are here shown in the second row. Note that this is an overall test where the haplotypes are analyzed combined. Measures of relative risk ratios for each stratum and haplotype are obtained by `haptable(res.POOxE)`.

## PoOxE Analysis on the X-Chromosome

PoOxE analyses on the X-chromosomes are carried out in a similar manner as for the autosomal markers, with the extension of three additional arguments. The argument `xchrom = T` enables analyses on X-linked markers. The `sex` argument indicates the data column containing the sex variable. In this example, we assume that a single allele in males has the same effect as a double allele dose in females. This corresponds to X-inactivation and is specified by `comb.sex = "double"`. However, if `comb.sex = "single"`, the effect of an allele in males is assumed to equal the effect of a single allele dose in females. PoOxE analyses on the X-chromosome can also be conducted for females only, as indicated by `comb.sex = "females"`.

```
res.PoOxE.xchrom <- haplinStrat(filename = "xchrom.dat",
                markers = 1, n.vars = 2, sex = 1,
                strata = 2, design = "triad",
                poo = T, xchrom = T,
                comb.sex = "double", response = "mult",
                reference = "ref.cat", use.missing = T)
```

## Power Calculations

The asymptotic power can be computed directly in Haplin by the function hapPowerAsymp. The function extracts the asymptotic standard error of the estimated log-parameter and then uses the properties of the non-centrality parameter of the chi-squared distribution.

If the minor allele at a dichotomous locus is associated with a two-fold risk only when inherited from the mother, the asymptotic power for 200 case-parent triads is calculated using the command:

```
power.PoO <- hapPowerAsymp(nall = 2,
        cases = c(mfc=200), haplo.freq = c(0.9,0.1),
        RRcm = c(1,2), RRcf = c(1,1), RRstar = c(1,1))
```

The number of alleles at each locus is given by the vector nall. The allele frequencies are specified by the argument haplo.freq, and the corresponding relative risks are indicated by RRcm and RRcf, depending on parental origin. The family design is given by the arguments cases and controls. The nominal significance level equals 0.05 unless otherwise specified.

The power of GxE effects might be examined by a command similar to

```
power.GxE <- hapPowerAsymp(nall = 2, n.strata = 2,
        cases = list(c(mfc=400),c(mfc=200)),
        haplo.freq = c(0.9,0.1),
        RR = list(c(1,1),c(1,2.5)), RRstar = c(1,1))
```

The argument n.strata indicates the number of strata. Here, the number of case-parent triads varies between the two exposure categories, and the least frequent allele is associated with disease only in the first stratum. The allele frequencies are the same in both strata. Extensions to several exposure levels are easily incorporated by modifying or expanding the appropriate arguments, e.g., `n.strata = 3`, `cases = list(c(mfc=400),c(mfc=200),c(mfc=100))` and `RR = list(c(1,1),c(1,2),c(1,3))`.

A power analysis for PoOxE interactions is achieved by combining the commands for PoO and GxE power calculations:

```
power.PoOxE <- hapPowerAsymp(nall = 2, n.strata = 2,
        cases = list(c(mfc=460),c(mfc=230)),
        haplo.freq = list(c(0.7,0.3),c(0.9,0.1)),
        RRcm = list(c(1,1),c(1,2.5)), RRcf = c(1,1),
        RRstar = c(1,1))


power.PoOxE
$haplo.power
  Haplotype RRcm.power RRcf.power RRcm_cf.power
1         1       0.94       0.05           0.8
2         2        ref        ref           ref
```

Here, there is only an effect of the maternally derived allele in the second stratum. The power to detect this change over strata is 94% (`RRcm.power`). The paternally derived allele has no effect in either stratum, so the corresponding power is 5%, i.e., equal to the nominal significance level. The actual PoOxE effect compares the two over strata, and thus has a somewhat lower power (`RRcm_cf.power`).

The statistical power for PoO, GxE and PoOxE interactions can also be computed through simulations. In Haplin, power simulations are carried out using a two-step procedure. First, `hapRun` is used to perform Haplin runs on simulated haplotype data, in which triad genotypes are generated from the multinomial distribution. The multinomial probabilities are calculated by listing all possible genotype combinations in the triad format. It then employs the sampling model (equation 1 in the main text), with appropriate adjustments to the relevant effect situations. The second step feeds the simulation results to hapPower, and the power is subsequently computed by calculating the fraction of p-values less than the nominal significance level.

Provided that the large-sample properties of the log-linear model hold, the asymptotic power should be comparable with that obtained from simulations. The asymptotic power for the PoOxE analysis can be verified by the following commands:

```
sim.power.PoOxE <- hapRun(nall = c(2), n.strata = 2,
        cases = list(c(mfc=460),c(mfc=230)),
        haplo.freq = list(c(0.7,0.3),c(0.9,0.1)),
        RRcm = list(c(1,1),c(1,2.5)), RRcf = c(1,1),
        RRstar = c(1,1), poo = T,
        hapfunc = "haplinStrat", response = "mult",
        n.sim = 1000, cpus = 4)
hapPower(sim.power.PoOxE)
```

The arguments of hapRun are similar to those of hapPowerAsymp with a few exceptions. In addition to the arguments RRcm and RRcf, poo must be set to true in order to test for PoO effects in hapRun. Also, one needs to specify which haplin function to run, the response model and the number of simulations. The argument cpus speeds up computations by allowing parallel processing.

Table S1: Values of $\omega^2$, the asymptotic variance of the log-parameter for **a)** a complete case-parent triad; and **b)** a complete case-mother dyad. The values are scaled to a sample of $n = 1$ triad or dyad, respectively.

| **a) Case-parent triad** | | | | | P | | |
|---|---|---|---|---|---|---|---|
| $RR_M$ | $RR_F$ | $RR_M RR_F$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
| 1 | 1 | 1 | 24.4 | 12.1 | 10.7 | 12.1 | 24.4 |
| 1 | 3/2 | 2/3 | 21.2 | 11.3 | 10.8 | 13.3 | 29.3 |
| 1 | 2 | 1/2 | 19.5 | 11.0 | 11.1 | 14.6 | 34.2 |
| 1 | 5/2 | 2/5 | 18.6 | 10.8 | 11.5 | 15.9 | 39.1 |
| 1 | 3 | 1/3 | 17.9 | 10.8 | 11.9 | 17.2 | 44.0 |
| 3/2 | 1 | 3/2 | 21.2 | 11.3 | 10.8 | 13.3 | 29.3 |
| 3/2 | 3/2 | 1 | 17.9 | 10.6 | 11.1 | 14.6 | 34.2 |
| 3/2 | 2 | 3/4 | 16.4 | 10.4 | 11.5 | 16.0 | 39.1 |
| 3/2 | 5/2 | 3/5 | 15.4 | 10.3 | 12.0 | 17.4 | 44.1 |
| 3/2 | 3 | 1/2 | 14.8 | 10.3 | 12.5 | 18.7 | 49.0 |
| 2 | 1 | 2 | 19.5 | 11.0 | 11.1 | 14.6 | 34.2 |
| 2 | 3/2 | 4/3 | 16.4 | 10.4 | 11.5 | 16.0 | 39.1 |
| 2 | 2 | 1 | 14.8 | 10.2 | 12.0 | 17.4 | 44.1 |
| 2 | 5/2 | 4/5 | 13.8 | 10.1 | 12.5 | 18.8 | 49.0 |
| 2 | 3 | 2/3 | 13.2 | 10.2 | 13.0 | 20.2 | 53.9 |
| 5/2 | 1 | 5/2 | 18.6 | 10.8 | 11.5 | 15.9 | 39.1 |
| 5/2 | 3/2 | 5/3 | 15.4 | 10.3 | 12.0 | 17.4 | 44.1 |
| 5/2 | 2 | 5/4 | 13.8 | 10.1 | 12.5 | 18.8 | 49.0 |
| 5/2 | 5/2 | 1 | 12.9 | 10.1 | 13.1 | 20.3 | 53.9 |
| 5/2 | 3 | 5/6 | 12.3 | 10.2 | 13.6 | 21.7 | 58.9 |
| 3 | 1 | 3 | 17.9 | 10.8 | 11.9 | 17.2 | 44.0 |
| 3 | 3/2 | 2 | 14.8 | 10.3 | 12.5 | 18.7 | 49.0 |
| 3 | 2 | 3/2 | 13.2 | 10.2 | 13.0 | 20.2 | 53.9 |
| 3 | 5/2 | 6/5 | 12.3 | 10.2 | 13.6 | 21.7 | 58.9 |
| 3 | 3 | 1 | 11.7 | 10.3 | 14.2 | 23.1 | 63.8 |

| **b) Case-mother dyad** | | | | | P | | |
|---|---|---|---|---|---|---|---|
| $RR_M$ | $RR_F$ | $RR_M RR_F$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
| 1 | 1 | 1 | 27.7 | 20.3 | 24.0 | 20.3 | 27.7 |
| 1 | 3/2 | 2/3 | 25.4 | 21.8 | 23.0 | 19.6 | 32.0 |
| 1 | 2 | 1/2 | 24.8 | 22.5 | 21.6 | 19.9 | 36.5 |
| 1 | 5/2 | 2/5 | 24.7 | 22.6 | 20.7 | 20.7 | 41.3 |
| 1 | 3 | 1/3 | 25.0 | 22.3 | 20.2 | 21.7 | 46.1 |
| 3/2 | 1 | 3/2 | 23.2 | 17.2 | 23.0 | 24.2 | 34.8 |
| 3/2 | 3/2 | 1 | 20.5 | 18.7 | 24.4 | 23.5 | 38.6 |
| 3/2 | 2 | 3/4 | 19.5 | 20.2 | 24.2 | 23.6 | 42.9 |
| 3/2 | 5/2 | 3/5 | 19.2 | 21.2 | 23.7 | 24.2 | 47.5 |
| 3/2 | 3 | 1/2 | 19.1 | 21.8 | 23.4 | 25.1 | 52.2 |
| 2 | 1 | 2 | 21.0 | 15.5 | 21.6 | 27.3 | 42.4 |
| 2 | 3/2 | 4/3 | 18.2 | 16.8 | 24.2 | 27.0 | 45.5 |
| 2 | 2 | 1 | 17.0 | 18.2 | 25.2 | 27.1 | 49.5 |
| 2 | 5/2 | 4/5 | 16.5 | 19.6 | 25.5 | 27.6 | 53.9 |
| 2 | 3 | 2/3 | 16.3 | 20.6 | 25.5 | 28.3 | 58.5 |
| 5/2 | 1 | 5/2 | 19.7 | 14.5 | 20.7 | 30.0 | 50.3 |
| 5/2 | 3/2 | 5/3 | 16.9 | 15.5 | 23.7 | 30.2 | 52.7 |
| 5/2 | 2 | 5/4 | 15.6 | 16.9 | 25.5 | 30.3 | 56.3 |
| 5/2 | 5/2 | 1 | 15.0 | 18.2 | 26.4 | 30.8 | 60.4 |
| 5/2 | 3 | 5/6 | 14.7 | 19.4 | 26.9 | 31.5 | 64.8 |
| 3 | 1 | 3 | 18.9 | 13.9 | 20.2 | 32.6 | 58.7 |
| 3 | 3/2 | 2 | 16.0 | 14.7 | 23.4 | 33.3 | 60.2 |
| 3 | 2 | 3/2 | 14.7 | 15.9 | 25.5 | 33.5 | 63.3 |
| 3 | 5/2 | 6/5 | 14.0 | 17.2 | 26.9 | 33.9 | 67.2 |
| 3 | 3 | 1 | 13.7 | 18.3 | 27.7 | 34.6 | 71.4 |

- $P$ is the frequency of the non-reference allele
- $RR_M$ and $RR_F$ are the relative risks depending on parental origin

Paper II

# Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls

Gjerdevik *et al. BMC Bioinformatics* (2019) 20:165
https://doi.org/10.1186/s12859-019-2727-3

BMC Bioinformatics

SOFTWARE Open Access

# Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls

Miriam Gjerdevik[1,2*], Astanand Jugessur[1,2,3], Øystein A. Haaland[1], Julia Romanowska[1,4], Rolv T. Lie[1,3], Heather J. Cordell[5] and Håkon K. Gjessing[1,3]

## Abstract

**Background:** Log-linear and multinomial modeling offer a flexible framework for genetic association analyses of offspring (child), parent-of-origin and maternal effects, based on genotype data from a variety of child-parent configurations. Although the calculation of statistical power or sample size is an important first step in the planning of any scientific study, there is currently a lack of software for genetic power calculations in family-based study designs. Here, we address this shortcoming through new implementations of power calculations in the **R** package Haplin, which is a flexible and robust software for genetic epidemiological analyses. Power calculations in Haplin can be performed analytically using the asymptotic variance-covariance structure of the parameter estimator, or else by a straightforward simulation approach. Haplin performs power calculations for child, parent-of-origin and maternal effects, as well as for gene-environment interactions. The power can be calculated for both single SNPs and haplotypes, either autosomal or X-linked. Moreover, Haplin enables power calculations for different child-parent configurations, including (but not limited to) case-parent triads, case-mother dyads, and case-parent triads in combination with unrelated control-parent triads.

**Results:** We compared the asymptotic power approximations to the power of analysis attained with Haplin. For external validation, the results were further compared to the power of analysis attained by the EMIM software using data simulations from Haplin. Consistency observed between Haplin and EMIM across various genetic scenarios confirms the computational accuracy of the inference methods used in both programs. The results also demonstrate that power calculations in Haplin are applicable to genetic association studies using either log-linear or multinomial modeling approaches.

**Conclusions:** Haplin provides a robust and reliable framework for power calculations in genetic association analyses for a wide range of genetic effects and etiologic scenarios, based on genotype data from a variety of child-parent configurations.

**Keywords:** Log-linear and multinomial models, Genome-wide association studies (GWAS), Statistical power estimation, Sample size estimation, Haplin, EMIM

*Correspondence: miriam.gjerdevik@uib.no
[1] Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
[2] Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway
Full list of author information is available at the end of the article

## Background

Statistical power or sample size analysis is an essential first step in the planning of any scientific study. Such analyses ensure that a study is capable of answering its stated research questions and are a prerequisite for optimal study design [1]. Furthermore, a power analysis is required in most research proposals. Statistical power calculations are particularly important in genome-wide association studies (GWAS) in order to maximize the scientific gains from the typically high genotyping and assay costs. Moreover, GWAS are often underpowered due to the large number of single-nucleotide polymorphisms (SNPs) being assessed, leading to issues of multiple testing. Most effect sizes reported from genetic association studies of complex traits are small [2–4], which further limits the power. The statistical power of a study affects the interpretation of the results. Low power may result in a high number of false negatives, and a power analysis might elucidate whether negative findings were the result of the study being underpowered.

Log-linear and multinomial modeling are closely related approaches that offer a flexible framework for genetic association analysis. Both approaches enable the estimation of genetic effects in addition to hypothesis testing. Beyond the standard case-control design, they are capable of incorporating child, parent-of-origin (PoO) and maternal effects based on genotype data from case-parent triads, as well as a range of other child-parent configurations. They can also handle incomplete triad data as well as independent controls. Moreover, the models are readily extended to haplotype analysis. Due to these appealing features, there has been much interest in the application of log-linear or multinomial models in genetic association studies [5–12], and the models are implemented in well-established software packages such as Haplin [10, 13] and EMIM (Estimation of Maternal, Imprinting and interaction effects using Multinomial modelling) [11, 12].

General-purpose software tools for statistical power and sample size analysis are not set up to handle the genetic study designs and effect estimates available from case-parent triads with unrelated controls. Although there are tools that offer power calculations for some genetic association studies, e.g., Quanto [14–16] and Genetic Power Calculator (GPC) [17], a comprehensive framework for power analysis based on the full triad design is lacking.

We propose a complete setup for power calculations tailored to binary disease traits, which we have implemented as a new module in the **R** package Haplin [10, 13]. In the new implementations, a power analysis can be performed based on the asymptotic variance-covariance structure of the parameter estimator or by a simulation procedure. The power for child, PoO, maternal, and gene-environment (GxE) effects are easily estimated. Haplin also enables power analyses for haplotypes, taking into account unknown SNP phase. The calculations can be performed for both autososomal and X-linked markers, and a variety of study designs can be accommodated.

Our paper is structured as follows. In the “Implementation” section, we first introduce the Haplin software and briefly present our new power calculation approaches. We then provide a short tutorial on power calculations for child, PoO and maternal effects, focusing on the use of asymptotic approximations. In the “Results” section, we illustrate our power calculations for a wide range of scenarios. We also compare our asymptotic power approximations to the powers attained by Haplin and EMIM in simulations, thus confirming the equivalent inference provided by log-linear and multinomial modeling. In Additional file 1, we derive the variance-covariance matrix underlying the asymptotic power calculations. Furthermore, because the Haplin framework includes numerous features for power analysis, we provide a more detailed and extensive tutorial, including power analysis for GxE interactions, in Additional file 2. In addition, we outline some of the possibilities for power calculations under different X-chromosome models, and we also show how the power calculations can be extended to haplotype analysis. Finally, we show the flexibility of our simulation approach, demonstrating different parameterization models and study designs.

## Implementation

Our power calculation tool has been added to the **R** package Haplin, which provides an extensive framework for genetic epidemiological analyses of binary traits. The new power calculation module has been integrated into the original setup for genetic association analysis in Haplin and is based on log-linear modeling, as previously described by Gjessing and Lie [10]. Haplin implements a full maximum-likelihood model for estimation and computes explicit estimates of relative risks with asymptotic standard errors and confidence intervals. It enables the estimation of child, PoO and maternal effects, as well as interactions between these genetic effects and categorical or ordinal exposure variables (i.e., GxE) [18, 19]. Haplin also incorporates analyses of X-linked markers in a straightforward manner, and different X-chromosome models may be fitted depending on the desired underlying assumptions [20–22]. In Haplin, the main unit of study is the case-parent triad, in which affected children and both of their biological parents are genotyped. However, the log-linear model can be extended to include independent control children or control triads in a hybrid design, under the “rare disease” assumption [23]. Note that unrelated controls are optional but not required, because “pseudo-controls” can be constructed from the non-transmitted parental alleles in case-parent triads [24–27]. The expectation maximization (EM) algorithm [28] is implemented in Haplin to account for unknown parental origin in ambiguous (uninformative) triads. Additionally, the EM algorithm accounts for missing information on certain individuals, such as when some triads are reduced to child-mother dyads due to missing data on the father. Although the fundamental model in Haplin relates to a single multi-allelic marker, it extends directly to haplotypes over multiple markers by statistically reconstructing haplotypes of unknown phase [10]. Furthermore, because the calculations can be performed in parallel, genome-wide association analyses are readily accommodated. The log-linear model in Haplin assumes Hardy-Weinberg equilibrium (HWE), Mendelian transmission and random mating. A detailed description of the underlying model is provided in several of our previous publications [10, 18, 29].

### Genetic effects and study designs

Within the Haplin framework, based on the log-linear modeling approach, we have developed a new and complete module for performing power calculations. The basic calculations relate to child, PoO and maternal effects, and our definitions of these genetic effects are provided in Table 1. The power depends on the underlying penetrance models, i.e., the probability of a child exhibiting the disease conditional on a particular genetic composition, which we define in Table 2. A variety of child-parent configurations are available for power analysis in Haplin, and a small selection of the possible study designs is shown in Fig. 1. We use the following abbreviations to describe the family designs. We let the letters c, m and f denote a child, mother and a father, respectively. Thus, mfc denotes a case-parent triad, and mc denotes a case-mother dyad. Moreover, mfc-mfc denotes the full hybrid design, whereas mc-mc denotes the hybrid design consisting of case-mother and unrelated control-mother dyads. The possible configurations in Haplin also include designs such as c-c (the standard case-control design), fc (case-father dyad), mfc-mc (case-parent triad with unrelated control-mother dyad) and mfc-mf (case-parent triad with unrelated control parents). The full list of supported study designs are provided on the Haplin website [13].

### Power calculations in Haplin

In this section, we demonstrate how to perform basic power calculations in Haplin, implemented in the function `hapPowerAsymp`. The power is computed analytically through asymptotic approximations, scaled to the appropriate sample size. We apply the asymptotic normal distribution of the log-scale parameter and use the chi-squared non-centrality parameter of the Wald test. The variance-covariance matrix is computed from a log-linear model which accounts for transmission ambiguities and missing data; its derivation is provided in Additional file 1. The theory underlying our asymptotic power calculations is outlined in more detail elsewhere [29].

In Haplin, the asymptotic power calculations are easy to perform. In general, one only needs to specify the study design and its sample size, the allele frequencies, and the type of genetic effect and its magnitude. Table 3 shows example Haplin commands for estimating the power for child, PoO and maternal effects. In all examples, we calculate the power for a diallelic SNP, using 500 case-parent triads. The study design is specified by the arguments `cases` and `controls`, using the notation from Fig. 1. Thus, 500 case-parent triads are specified by the argument `cases=c(mfc=500)`, whereas 500 case-mother dyads would be specified by `cases=c(mc=500)`. A hybrid design consisting of 200 case-mothers dyads and 500 control-parent triads would be expressed by the combination `cases=c(mc=200)` and `controls=c(mfc=500)`.

The genetic effects are determined by the choice of relative risk parameter(s), which also specifies the effect

**Table 1** Genetic effects

| Effects | Description |
|---|---|
| Child | A variant allele may increase the risk of a disease only when carried by an individual himself/herself. We refer to this as a “child effect” since it is frequently estimated from the offspring in a case-parent triad. However, the individual referred to as a child might be of any age, depending on the phenotype of interest, and the same effect can also be estimated in case-control studies. |
| Parent-of-origin (PoO) | A PoO effect occurs if the effect of a variant allele in the child depends on whether it is inherited from the mother or the father. In statistical terms, we define a PoO effect as the interaction effect $\mathrm{RRR} = \mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$, which is a measure of the risk increase (or decrease) associated with allele $A_j$, when derived from the mother as opposed to the father. In contrast, regular child-effect analyses assume that the effect of an allele in the child is independent of parental origin. Note that genomic imprinting (an epigenetic phenomenon where one of the inherited parental alleles is expressed whereas the other is silenced) may cause PoO effects [32]. |
| Maternal | A mother’s genotype may influence fetal development directly, for example through maternal metabolic factors operating in utero [33], and may affect health throughout life [34]. A maternal effect occurs when a variant allele carried by the mother increases the risk of disease in her child, regardless of whether or not the allele has been transferred to the child [35]. This is distinct from child and PoO effects, in which we measure the effect of alleles in the child himself/herself. Because these underlying genetic mechanisms lead to entirely different biological interpretations, distinguishing between the genetic effects is particularly important in advancing the understanding of the etiology underlying a complex disease [11, 36, 37]. |

**Table 2** Parameterization of penetrances

| Effects | Parameterization of penetrances | |
|---|---|---|
| Child | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl}$ | (1) |
| Parent-of-origin (PoO) | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl}$ | (2) |
| Child and maternal | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ | (3) |
| PoO and maternal | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ | (4) |

*B* is the baseline risk level, typically associated with the (more common) reference allele; $\mathrm{RR}_j$ is the risk increase associated with allele $A_j$, relative to *B*; $\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ are the relative risks associated with allele $A_j$, depending on whether the allele is transmitted from the mother or the father; the double-dose parameter $\mathrm{RR}^*_{jl}$ measures the deviation from what would be expected in a multiplicative dose-response relationship, i.e., $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and $\mathrm{RR}^*_{jl} = 1$ when $j \neq l$; $\mathrm{RR}^{(M)}_i$ is the relative risk associated with allele $A_i$ carried by the mother, and $\mathrm{RR}^{(M)*}_{ij}$ is the maternal double-dose parameter, interpreted analogously to $\mathrm{RR}^*_{ij}$. To ensure that the model is not overparameterized, we set $\mathrm{RR} = 1$ for the reference allele

sizes. Corresponding to the parameterization model in Eq. (1) (defined in Table 2), a child effect is specified by the relative risk argument `RR` (Table 3a). Allele frequencies are specified by the argument `haplo.freq`. Note that the order and length of the specified relative risk parameter vectors should always match the corresponding allele frequencies. All examples assume a minor allele frequency (MAF) of 0.2. Thus, from Table 3a we see that the power is 88% when the less frequent allele at a diallelic marker is associated with a relative risk of 1.4, as expressed by the combination of allele frequencies `haplo.freq=c(0.8,0.2)` and relative risks `RR=c(1,1.4)`. By default, the more frequent allele is chosen as reference (Table 3a, first row of the Haplin output).

As illustrated in Table 3b, the power to detect a PoO effect is computed by replacing the argument `RR` by the two relative risk arguments `RRcm` and `RRcf`, denoting parental origin `m` (mother) and `f` (father). Both $\mathrm{RR}_M$ and $\mathrm{RR}_F$ are estimated freely, and individual tests for the null hypotheses $\mathrm{RR}_M = 1$ and $\mathrm{RR}_F = 1$ are constructed. The corresponding power estimates are denoted by `RRcm.power` and `RRcf.power`, respectively. In addition, we are interested in testing the actual PoO effect, estimated by comparing the maternally and paternally derived effects by the ratio $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F$. The null hypothesis of $\mathrm{RRR} = \mathrm{RR}_M/\mathrm{RR}_F = 1$ means no PoO effect, and the power to detect the PoO effect is output as `RRcm_cf.power`, here estimated to be 48% when `RRcm = c(1,2)` and `RRcf = c(1,1.5)`. For more details on PoO testing and its relationship to imprinting, see Gjerdevik et al. [29].

Since children and their mothers have an allele in common, a maternal effect might be statistically confounded with a child or a PoO effect. Corresponding to the parameterization models in Eq. (3) and (4) (Table 2), the power of a maternal effect can be analyzed jointly with that of a child effect or a PoO effect by adding the relative risk argument `RR.mat` to the original child or PoO model (Table 3c and d). The resulting power estimates control for the possible confounding of these effects with one another. When adjusting for the maternal effect in Table 3c, the power to detect the child effect is 90%. Conversely, when adjusting for the child effect, the power to

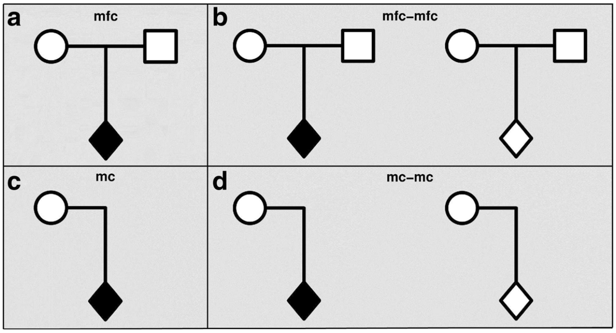

**Fig. 1** A selection of designs for genetic association studies: **a** Case-parent triad (mfc); **b** Case-parent triad with independent control-parent triad (mfc-mfc); **c** Case-mother dyad (mc); **d** Case-mother dyad with independent control-mother dyad (mc-mc)

**Table 3** Examples of asymptotic power calculations in Haplin

| Effects | Haplin commands | Output |
|---|---|---|
| a) Child | `hapPowerAsymp(cases = c(mfc=500),`<br>`haplo.freq = c(0.8,0.2),`<br>`RR = c(1,1.4))` | `$haplo.power`<br>`Haplotype RR.power`<br>`1 ref`<br>`2 0.88` |
| b) PoO | `hapPowerAsymp(cases = c(mfc=500),`<br>`haplo.freq = c(0.8,0.2),`<br>`RRcm = c(1,2), RRcf = c(1,1.5))` | `$haplo.power`<br>`Haplotype RRcm.power RRcf.power RRcm_cf.power`<br>`1 ref ref ref`<br>`2 1 0.87 0.48` |
| c) Child and maternal | `hapPowerAsymp(cases = c(mfc=500),`<br>`haplo.freq = c(0.8,0.2),`<br>`RR = c(1,1.4), RR.mat = c(1,1.2))` | `$haplo.power`<br>`Haplotype RR.power RRm.power`<br>`1 ref ref`<br>`2 0.9 0.42` |
| d) PoO and maternal | `hapPowerAsymp(cases = c(mfc=500),`<br>`haplo.freq = c(0.8,0.2),`<br>`RRcm = c(1,2), RRcf = c(1,1.5),`<br>`RR.mat = c(1,1.2))` | `$haplo.power`<br>`Haplotype RRcm.power RRcf.power RRcm_cf.power RRm.power`<br>`1 ref ref ref ref`<br>`2 0.99 0.65 0.2 0.17` |

The power is calculated for a diallelic SNP, using 500 case-parent triads (`cases = c(mfc=500))`, and a MAF of 0.2 (`haplo.freq = c(0.8,0.2)`). The argument `RR` specifies the relative risk associated with the child effect, whereas the power to detect a PoO effect is calculated by replacing `RR` by the two relative risk arguments `RRcm` and `RRcf`, which refer to the parental origin of the allele carried by the child. Maternal effects can be included by adding the maternal relative risk parameter `RR.mat` to the original child or PoO command. Note that the order of alleles to which the relative risk parameters refer corresponds to the order used for the haplotype frequencies. Here, the less frequent allele is set as the risk allele and the more frequent allele is used as reference. The nominal significance level defaults to 0.05, but other values can be specified by the argument `alpha` detect the maternal effect is 42%. The example in Table 3d, involving joint PoO and maternal effects, has a similar interpretation.

In Table 3, the nominal significance level defaults to 5%. However, other values can be specified by using the argument alpha. The current implementation of hapPowerAsymp does not allow deviations from the multiplicative dose-response assumption. Thus, the double-dose parameters RR* and $\text{RR}^{(M)*}$ (Eq. 1-4 in Table 2) are equal to 1 and do not need to be specified in the Haplin command. However, we expect future versions of hapPowerAsymp to handle power calculations for separate single- and double-dose effects.

### Power simulations in Haplin

Haplin also includes an extensive setup for power calculation through simulations. Simulation approaches are robust ways of checking software implementations, attained power, and attained significance level. They are particularly useful for small to moderately sized datasets, in which the asymptotic properties of the log-linear model might not hold true. In these situations, the extent and direction of the possible bias can best be assessed using simulations. In Haplin, power simulations are carried out using a two-step approach, by applying the functions hapRun and hapPower. First, hapRun simulates haplotype data, in which triad genotypes are generated from the multinomial distribution. The multinomial probabilities are computed by listing all possible genotype combinations in the triad format and then applying the sampling model described in Gjessing and Lie [10]. hapRun then performs Haplin runs, i.e., statistical inference, on the simulated data. To speed up these calculations, hapRun allows for parallel processing. In the second step, the simulation results from hapRun are submitted to hapPower, which computes the power by calculating the fraction of p-values less than the nominal significance level.

Clearly, the asymptotic power approximation is much more time-efficient than brute-force simulations; in its current implementation, however, it is somewhat more restricted. The simulation approach is completely general; it enables power calculations for a wider range of parameterization models, such as deviations from the multiplicative dose-response assumption. The simulation approach also handles a wider array of child-parent configurations and allows for missing individuals to be generated at random. Examples and relevant Haplin commands are provided in Additional file 2.

## Results

### Examples of asymptotic power calculations

We illustrate the use of our power function hapPowerAsymp by plotting power curves for different scenarios, as shown in Fig. 2. Power calculations for child effects are shown in panels **a** and **b**, and power calculations for PoO effects are shown in panels **c** and **d**. For the PoO effects, we set $\text{RR}_F = 1$, so that the value of $\text{RRR} = \text{RR}_M/\text{RR}_F$ is equal to the value of $\text{RR}_M$. In the left panels (**a** and **c**), we used varying numbers of case-parent triads and a MAF of 0.2. In the right panels (**b** and **d**), the power was calculated using varying MAFs and a total of 500 case-parent triads. We used a nominal significance level of 5% throughout.

In all panels, the green, solid line represents scenarios in which 500 case-parent triads and a MAF of 0.2 were used. For child effects, we have 80% power to detect an RR of 1.35. However, using 250 case-parent triads, the corresponding power decreases to 51% (panel **a**). Moreover, with 500 case-parent triads and a MAF of 0.1, the power to detect an RR of 1.35 is 57% (panel **b**). The PoO analysis can be viewed as a statistical interaction. Compared with the child-effect analysis, a higher sample size is therefore required for the PoO analysis to reach the same statistical power for a similar effect size. Approximately 1200 case-parent triads are needed to detect an RRR of 1.35 with 80% power and a MAF of 0.2 (panel **c**). With 500 case-parent triads and a MAF of 0.2, we have approximately 80% power to detect an RRR of 1.6. Using a MAF of 0.1, the corresponding power is 64% (panel **d**).

Note that sample size and power are directly related measures. For given relative risks, power curves similar to Fig. 2 can be made with sample size on the x-axis.

### Comparison of the asymptotic power approximations to the simulated power in Haplin and EMIM

Similar to Haplin, the command line software PREMIM and EMIM are easy-to-use tools for the estimation of child, PoO and maternal effects based on genotype data from a number of different study designs [11, 12]. PREMIM generates required input files for EMIM by extracting the required genotype data from standard-format pedigree data (PLINK) files [30], and EMIM performs the subsequent statistical analyses. PREMIM and EMIM are written in C++ and FORTRAN 77, respectively, and are therefore considerably faster than **R** implementations. EMIM allows a variety of different parameterization models, which makes it an appealing software for power comparisons with Haplin. Because EMIM uses multinomial modeling, its inference should be similar to that of Haplin [31]. However, to account for unknown parental origin in ambiguous (uninformative) triads or dyads, EMIM maximizes the multinomial likelihood directly (via a direct search algorithm), whereas Haplin maximizes the likelihood using the EM algorithm.

We compared the asymptotic power calculations in Haplin to the power attained by Haplin and EMIM in data simulations. The asymptotic power was computed

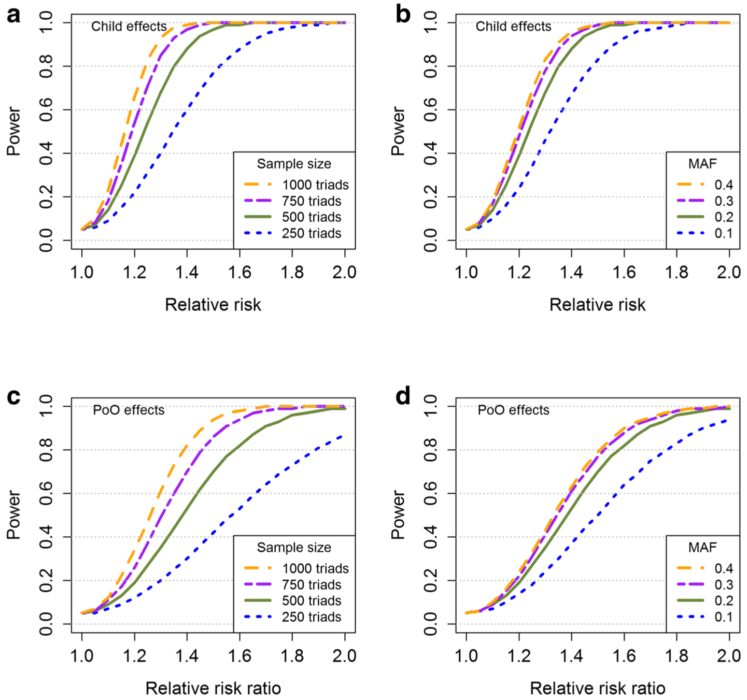

**Fig. 2** Power analysis using the Haplin function `hapPowerAsymp`. **a** Child effects for varying numbers of case-parent triads, using a MAF of 0.2; **b** Child effects for varying values of MAFs, using a total of 500 case-parent triads; **c** PoO effects for varying numbers of case-parent triads, using a MAF of 0.2; **d** PoO effects for varying values of MAFs, using a total of 500 case-parent triads. For the PoO effects, $RR_F = 1$, so that the value of $RR_M/RR_F$ is equal to $RR_M$. A nominal significance level of 0.05 was used throughout. The power was calculated at relative risks/relative risk ratios of $1, 1.05, 1.10, \ldots, 2$. Intermediate values correspond to line segments joining two adjacent points

using the function `hapPowerAsymp`, whereas the simulated power in Haplin was calculated using `hapRun` and `hapPower`. EMIM performs genetic association analyses, but corresponding power calculations are not implemented. To calculate the power attained by EMIM, we first used the Haplin function `hapSim` to simulate the genotype data. The data was then converted to the standard PLINK-format files, which were subsequently fed into PREMIM and EMIM. Given that the power calculations in Haplin are based on the Wald test, we also used the Wald test for inference in EMIM. Lastly, we calculated the fraction of p-values less than the nominal significance level. We analyzed child, PoO and maternal effects employing the parameterizations presented in Table 2, assuming a multiplicative dose-response model. We simulated data for a variety of child-parent configurations (mfc, mc, mfc-mfc, mc-mc), with effect sizes ranging between 1.0 and 2.0, and a MAF of 0.2. We based the power comparisons on 500 case families in each design, i.e., 500 case-mother dyads or 500 case-parent triads, reflecting that the number of case children available is often a constraint when designing a study. For the hybrid designs, we added an equal number of unrelated control families. The simulations were based on 10,000 replicates of data for a single SNP, and we used a nominal significance level of 0.05. HWE and random mating were assumed throughout.

The results are shown in Fig. 3. Child effects are displayed in panels **a** and **b**, and PoO effects are displayed in panels **c** and **d**, with panels **b** and **d** showing the results obtained when the child and PoO effects were calculated while adjusting for possible maternal effects (even though, in the simulation model, we did not assume maternal effects, i.e., we set $RR^{(M)} = 1$). For the PoO effects, we set $RR_F = 1$, so that the value of $RR_M/RR_F$ is equal to the value of $RR_M$. Panels **e** and **f** show the power to detect maternal effects, while adjusting for possible child or PoO effects (simulated under models where no such child or PoO effects existed, i.e., $RR^{(M)} > 1$ and $RR = 1$, and $RR^{(M)} > 1$ and $RR_M = RR_F = 1$, respectively).

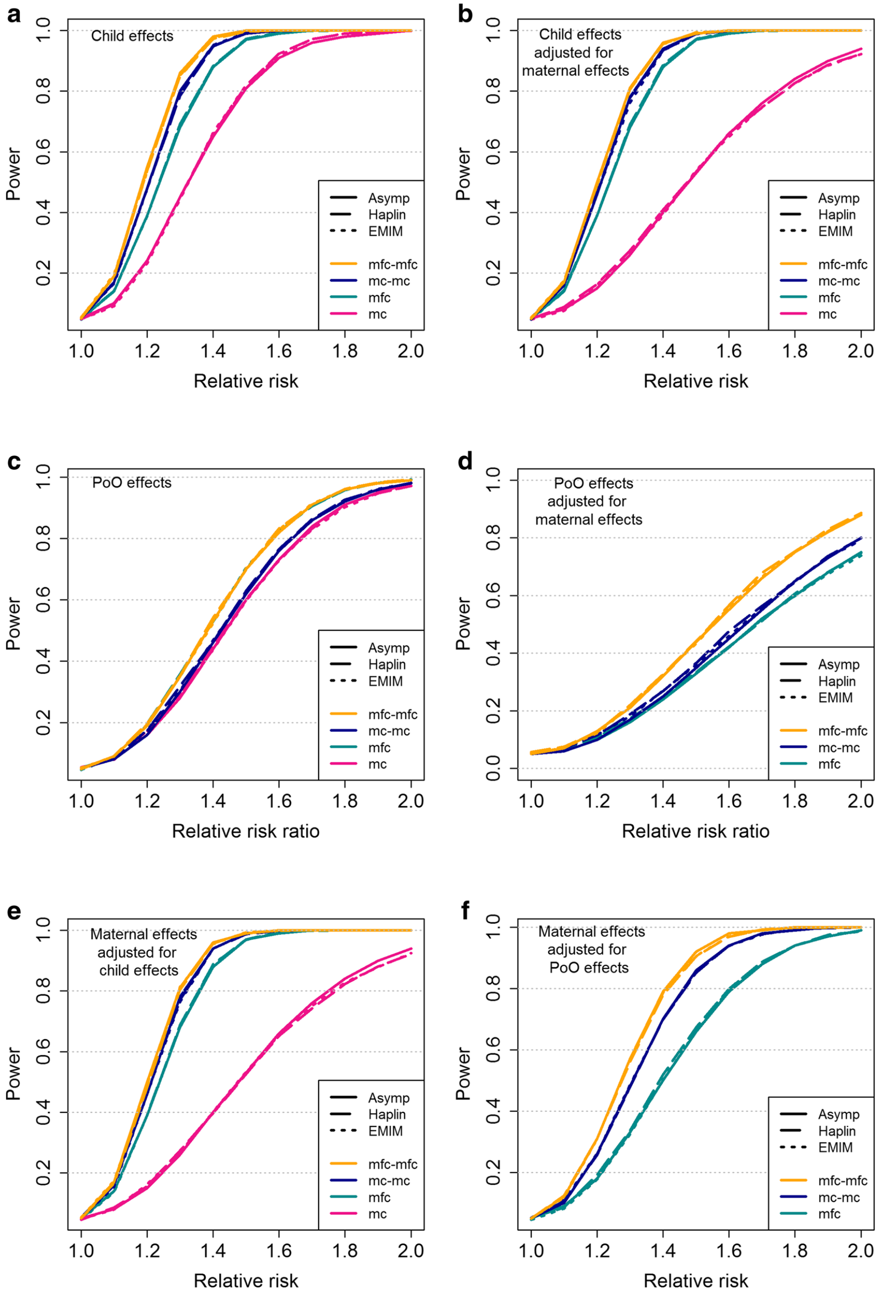

**Fig. 3** (See legend on next page.)

**Fig. 3** (See figure on previous page.)
Comparison of the asymptotic power calculations with the power attained by Haplin and EMIM in data simulations. The power was calculated for different child-parent configurations, assuming a MAF of 0.2 and a nominal significance level of 0.05. The results were based on 500 case families and, when applicable, 500 unrelated control families. All simulations were based on 10,000 replicates of data for a single SNP. **Asymp**: Power calculations in Haplin, based on asymptotic approximations (Haplin function `hapPowerAsymp`); **Haplin**: Power calculations in Haplin, based on data simulations. The power is the proportion of tests rejected by Haplin (Haplin functions `hapRun` and `hapPower`); **EMIM**: Power calculations based on data simulations in Haplin (Haplin function `hapSim`). The power is the proportion of tests rejected by EMIM. **a** Child effects (RR $> 1$); **b** Child effects, adjusting for maternal effects (RR $> 1$ and RR$^{(M)} = 1$); **c** PoO effects (RR$_M$/RR$_F > 1$ and RR$_F = 1$); **d** PoO effects, adjusting for maternal effects (RR$_M$/RR$_F > 1$ and RR$_F =$ RR$^{(M)} = 1$); **e** Maternal effects, adjusting for child effects (RR$^{(M)} > 1$ and RR $= 1$); **f** Maternal effects, adjusting for PoO effects (RR$^{(M)} > 1$ and RR$_M =$ RR$_F = 1$). The power was calculated at relative risks/relative risk ratios of 1, 1.1, 1.2, . . . , 2. Intermediate values correspond to line segments joining two adjacent points. Note that for all study designs, the power was calculated based on asymptotic approximations in Haplin, as well as simulations where both Haplin and EMIM were used to analyze the genetic data. The lines for Asymp, Haplin and EMIM are nearly overlapping, demonstrating consistent results

Note that panels **b** and **e** are equivalent because the power to detect a given child or maternal effect is identical when adjusting for possible confounding of the effects with one another. However, this symmetry depends on the study design and will not necessarily hold if case-mothers are unavailable for genotyping (results not shown). PoO effects are essentially estimated in case families, by contrasting the frequencies of alleles transmitted from mother to child with those of alleles transmitted from father to child. Thus, unrelated control families do not add extra power to the case-parent triad design, as can be seen from the overlapping results of the mfc and mfc-mfc designs in panel **c**. Note that we excluded the mc design from the joint PoO and maternal effect-analyses (panels **d** and **f**) because the penetrance model in Eq. (4) (Table 2) would become overparameterized. Overall, Fig. 3 shows that the results are highly consistent between the asymptotic power approximations and the simulated power in Haplin and EMIM, demonstrating that the asymptotic power function performs well when the asymptotic properties underlying the log-linear model hold true. Furthermore, the consistency between Haplin and EMIM across a wide spectrum of genetic scenarios confirms the computational accuracy of the inference methods used in both programs. Altogether, the results indicate that Haplin provides a robust and reliable framework for power calculations in genetic association studies when the genetic analyses are based on either log-linear or multinomial modeling.

## Conclusions

To our knowledge, a comprehensive software for power analysis based on the full triad design has been lacking. Here, we have developed and showcased extensive, new and easy-to-use functionalities for statistical power analyses based on log-linear modeling, incorporated in the **R** package Haplin. In Haplin, power analysis can be carried out analytically using the asymptotic variance-covariance structure of the parameter estimator, or, by a straightforward simulation procedure. The two approaches for power calculations complement each other, balancing time efficiency against generality. Haplin enables power calculations to be performed for child, PoO, maternal and GxE effects, based on genotype data from a variety of family-based study designs. An inherent strength of the Haplin framework is its ability to compute power for both single SNPs and haplotypes, either autosomal or X-linked. We plan to continue to expand the present framework for power analysis, adding new features for power calculations as additional methods for genetic association analysis are developed and incorporated into the Haplin software.

To facilitate power analysis in Haplin, we have provided relevant example commands in Table 3. In addition, an extended tutorial is provided in Additional file 2, demonstrating power analysis for GxE interactions, X-linked models and haplotype effects, as well as our simulation functions `hapRun` and `hapPower`. Researchers can easily apply our functions using arguments and parameter values relevant to their own data.

The standard Haplin implementation assumes haplotype-frequency parameters under HWE instead of a model with all mating-type parameters [5, 6]. This improves power and facilitates haplotype reconstruction. The triad design itself protects against population stratification, but some of that benefit is lost if HWE is not fulfilled. However, top hits from a GWAS analysis can be checked retrospectively for HWE. As for power calculations, a full set of mating-type frequencies will seldom be available prior to study start, and a HWE assumption simplifies the calculations.

We conducted a thorough comparison of the asymptotic approximation approach with the power attained by Haplin and EMIM in data simulations. Child, PoO and maternal effects were assessed. The concordant results obtained confirm the computational accuracy of the inference methods used in both programs. They also demonstrate that power calculations in Haplin are applicable to genetic association studies analyzed by either log-linear or multinomial modeling approaches. Thus, Haplin provides a robust and reliable framework for power calculations in genetic association analyses for various genetic effects and etiologic scenarios, based on genotype data from a wide range of different child-parent configurations.

## Availability and requirements

**Project name:** Haplin
**Project home page:** https://people.uib.no/gjessing/genetics/software/haplin
**Operating system(s)**: Platform independent
**Programming language:** Haplin is implemented as a standard package in the statistical software **R**. It is available from the official **R** package archive, CRAN (https://cran.r-project.org).
**Other requirements:** None
**License:** GPL (>= 2)
**Any restrictions to use by non-academics:** None
Information on EMIM and PREMIM is available from https://www.staff.ncl.ac.uk/richard.howey/emim.

## Additional files

**Additional file 1:** An asymptotic approximation of $\Sigma$. (PDF 184 kb)
**Additional file 2:** Power and sample size calculations in Haplin. (PDF 369 kb)

### Abbreviations

EM algorithm: The expectation maximization algorithm; GxE: Gene-environment interaction; GWAS: Genome-wide association study; HWE: Hardy-weinberg equilibrium; MAF: Minor allele frequency; mc: The case-mother dyad design; mc-mc: Case-mother dyads with unrelated control-mother dyads; mfc: The case-parent triad design; mfc-mfc: Case-parent triads with unrelated control-parent triads; PoO effect: Parent-of-origin effect; RR: Relative risk; RRR: Relative risk ratio; SNP: Single-nucleotide polymorphism

### Acknowledgements

Not applicable.

### Funding

This work has been supported by the Wellcome Trust (Grant 102858/Z/13/Z), by the Bergen Medical Research Foundation (BMFS) (Grant 807191), and by the Research Council of Norway (RCN) through Biobank Norway (Grant 245464/F50) and the Centres of Excellence funding scheme (Grant 262700). The funding bodies played no role in the design of the study, analysis or interpretation of data, nor in writing the manuscript.

### Availability of data and materials

The attained significance level and power were assessed using data simulations, available through the Haplin functions `hapSim` and `hapRun` (see https://people.uib.no/gjessing/genetics/software/haplin). The power simulation procedure in Haplin has been described in the main article, as well as in Additional file 2, and the source code is available from CRAN (https://cran.r-project.org).

### Authors' contributions

MG developed the power calculation tools in Haplin, performed computer simulations, conceived and planned the experiments and drafted the manuscript. AJ, ØAH, JR and RTL helped develop the concepts and revised the manuscript. JR has also contributed to the recent developments of Haplin. HJC developed the EMIM software, conceived and planned the experiments and revised the manuscript. HKG developed the Haplin software, conceived and planned the experiments and revised the manuscript. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

[1]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway. [2]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway. [3]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. [4]Computational Biology Unit, University of Bergen, Bergen, Norway. [5]Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, UK.

Received: 12 October 2018 Accepted: 13 March 2019
Published online: 02 April 2019

### References

1. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15(5):335–46.
2. Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol. 2006;164(7):609–14.
3. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007;39(1):17–23.
4. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10):1520–8.
5. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet. 1998;62(4):969–78.
6. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". Am J Epidemiol. 1998;148(9):893–901.
7. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet. 1999;65(1):229–35.
8. Umbach DM, Weinberg CR. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet. 2000;66(1):251–61.
9. Sinsheimer JS, Palmer CGS, Woodward JA. Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. Genet Epidemiol. 2003;24(1):1–13.
10. Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. Ann Hum Genet. 2006;70(3):382–96.
11. Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. Genet Epidemiol. 2011;35(1):19–45.
12. Howey R, Cordell HJ. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. BMC Bioinformatics. 2012;13:149.
13. Gjessing HK. Haplin: analyzing case-parent triad and/or case-control data with SNP haplotypes. 2018. R package version 7.0.0. Available from: https://people.uib.no/gjessing/genetics/software/haplin.

14. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155(5):478–84.
15. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med. 2002;21(1):35–50.
16. Gauderman WJ. Candidate gene association analysis for a quantitative trait, using parent-offspring trios. Genet Epidemiol. 2003;25(4):327–38.
17. Purcell S, Cherny SS, Sham PC. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003;19(1):149–50.
18. Skare Ø, Jugessur A, Lie RT, Wilcox AJ, Murray JC, Lunde A, et al. Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. Ann Hum Genet. 2012;76(3):221–36.
19. Haaland ØA, Lie RT, Romanowska J, Gjerdevik M, Gjessing HK, Jugessur A. A genome-wide search for gene-environment effects in isolated cleft lip with or without cleft palate triads points to an interaction between maternal periconceptional vitamin use and variants in *ESRRG*. Front Genet. 2018;9:60.
20. Jugessur A, Skare Ø, Lie RT, Wilcox AJ, Christensen K, Christiansen L, et al. X-linked genes and risk of orofacial clefts: evidence from two population-based studies in Scandinavia. PLoS ONE. 2012;7(6):e39240.
21. Skare Ø, Gjessing HK, Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, et al. A new approach to chromosome-wide analysis of X-linked markers identifies new associations in Asian and European case-parent triads of orofacial clefts. PLoS ONE. 2017;12(9):e0183772.
22. Skare Ø, Lie RT, Haaland ØA, Gjerdevik M, Romanowska J, Gjessing HK, et al. Analysis of parent-of-origin effects on the X chromosome in Asian and European orofacial cleft triads identifies associations with *DMD, FGF13, EGFL6*, and additional loci at Xp22.2. Front Genet. 2018;9:25.
23. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. Am J Hum Genet. 2005;77(4):627–36.
24. Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet. 1993;52(6):1085–93.
25. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet. 1993;53(5):1114–26.
26. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol. 2004;26(3):167–85.
27. Cordell HJ. Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. Genet Epidemiol. 2004;26(3):186–205.
28. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol). 1977;39(1):1–38.
29. Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. Ann Hum Genet. 2018;82(2):60–73.
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
31. Baker SG. The multinomial-poisson transformation. J R Stat Soc Ser D (Stat). 1994;43(4):495–504.
32. Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. Nat Rev Genet. 2013;14(9): 609–17.
33. Guilmatre A, Sharp AJ. Parent of origin effects. Clin Genet. 2012;81(3): 201–9.
34. Kong A, Thorleifsson G, Frigge ML, Vilhjalmsson BJ, Young AI, Thorgeirsson TE, et al. The nature of nurture: effects of parental genotypes. Science. 2018;359(6374):424–8.
35. Connolly S, Heron EA. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. Brief Bioinform. 2015;16(3): 429–48.
36. Hager R, Cheverud JM, Wolf JB. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. Genetics. 2008;178(3):1755–62.
37. McGinnis R, Steinthorsdottir V, Williams NO, Thorleifsson G, Shooter S, Hjartardottir S, et al. Variants in the fetal genome near *FLT1* are associated with risk of preeclampsia. Nat Genet. 2017;49(8):1255–60.

# Additional file 1 — An asymptotic approximation of $\Sigma$

**Likelihood model** Assume a locus has $l$ different alleles. Typically, $l = 2$ at a single SNP. With $k$ SNPs, there are $l = 2^k$ different possible haplotypes, each considered an allele at the locus, assuming no recombination between SNPs. For a mother-father-child triad, the genotype of the triad can be written as $(A_{M1}A_{M2}, A_{F1}A_{F2}, A_{C1}A_{C2})$, where $A_M$ denotes the maternal alleles, $A_F$ denotes the paternal alleles, and $A_C$ denotes the alleles of the child. Assume that the parental alleles are ordered in such a way that the second allele is transmitted to the child; i.e., we have $A_{C1} = A_{M2}$ and $A_{C2} = A_{F2}$. This permits a more compact notation with the full triad as an ordered quadruplet $(A_{M1}, A_{M2}, A_{F1}, A_{F2})$.

To list all possible triad genotypes, we construct a 4-column matrix $\boldsymbol{G}$ with one column for each of the parental alleles, including all possible allele combinations. For instance, for a diallelic SNP with alleles 1 and 2,

$$\boldsymbol{G} = \begin{array}{c} \begin{array}{cccc} A_{M1} & A_{M2} & A_{F1} & A_{F2} \end{array} \\ \left( \begin{array}{cccc} 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 2 & 1 & 2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 2 & 2 & 2 & 2 \end{array} \right) \end{array} .$$

The matrix $\boldsymbol{G}$ has dimensions $q \times 4$, where $q = l^4$. In particular, $q = 2^{4k}$ when considering a locus with $k$ diallelic SNPs, where the alleles are the $2^k$ possible haplotypes at the locus.

Assuming the full genotype of all triads could be observed, the log-linear model assumes that the number of triads $\boldsymbol{n} = [n_1, \ldots, n_q]$, corresponding to the rows of $\boldsymbol{G}$, can be described by independent Poisson distributions, where

$$\boldsymbol{m} = \exp(\boldsymbol{X\beta})$$

is a $q \times 1$ vector of the expected number of triads in each row, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, $\boldsymbol{X}$ is a $q \times p$ design matrix (described in more detail below), and the exponential function is computed elementwise. We assume that $\boldsymbol{1} \in \text{colspace}(\boldsymbol{X})$, where $\boldsymbol{1} = [1, \ldots, 1]^T$. If $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator derived from this model, and $\hat{\boldsymbol{m}} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})$, the condition

$$\hat{\boldsymbol{m}}^T \boldsymbol{1} = \boldsymbol{n}^T \boldsymbol{1}$$

holds [1, Chapter 10], i.e., the sum of the expected number of triads is equal to the sample size $N = \boldsymbol{n}^T\boldsymbol{1}$. Let $m_{\boldsymbol{\cdot}} = \boldsymbol{m}^T\boldsymbol{1}$, and define $\boldsymbol{p} = \boldsymbol{m}/m_{\boldsymbol{\cdot}}$, i.e., the cell probabilities.

Each row in $\boldsymbol{G}$ corresponds to a fully observed triad genotype, that is, a triad can be associated with a specific row of $\boldsymbol{G}$ only if the alleles of the mother, father, and child are all fully known. From observed data, however, one will often obtain triads where, for instance, the genotypes of the father is lacking. Also, since the alleles at the locus will typically consist of haplotypes derived from a sequence of SNPs, the unknown phase of the SNPs will represent an ambiguity regarding the triad alleles. For any observed triad $j$, we define $\boldsymbol{a}_j$ to be the $q \times 1$ "ambiguity vector" for triad $j$. To determine $\boldsymbol{a}_j$ for a given triad, we first identify all rows of $\boldsymbol{G}$ that are compatible with the observed genotype of the triad. For instance, at a SNP, if we observe a mother with genotype (1 2), a child with the genotype (2 2), and the father is missing, the full triad could be either (1 2, 1 2, 2 2) or (1 2, 2 2, 2 2), which correspond to rows 11 and 15, respectively, in the $\boldsymbol{G}$ matrix. The ambiguity vector $\boldsymbol{a}_j$ is then a vector with ones at positions 11 and 15, and zeros otherwise. Similarly, with two or more SNPs, unknown haplotype phase introduces ambiguities which are incorporated in the ambiguity vector $\boldsymbol{a}_j$.

Let $\mathcal{A}$ be the set of all possible ambiguity vectors, i.e., those corresponding to all observed genotypes. Note that $\boldsymbol{a_j}$ is a many-to-one mapping from the rows of $\boldsymbol{G}$ into $\mathcal{A}$, and thus $P(\boldsymbol{a_j} = \boldsymbol{a}) = \boldsymbol{a}^T\boldsymbol{p}$, i.e., the sum over all row probabilities compatible with the observed genotype.

**Design matrix** Using $\boldsymbol{G}$ as the starting point, the corresponding $q \times p$ design matrix $\boldsymbol{X}$ for a log-linear model can be derived, including columns for estimating allele frequencies, child allele dose effects, etc. The form of $\boldsymbol{X}$ will depend on what model is being estimated in a given instance. For example, to estimate the child relative risk $\mathrm{RR}_2$ associated with allele 2, we first create two dummy vectors $1_M$ and $1_F$. The dummy $1_M$ is set to one when the $A_{M2}$ column of $\boldsymbol{G}$ is equal to 2, and zero otherwise. Similarly, $1_F$ is set to one when the $A_{F2}$ column is 2, and zero otherwise. That is, $1_M$ and $1_F$ indicate whether the child inherited allele 2 from the mother and/or the father, respectively. The design matrix $\boldsymbol{X}$ should then contain a column equal to $1_M + 1_F$, and $\hat{\mathrm{RR}}_2 = \exp(\hat{\beta})$, where $\hat{\beta}$ is the estimated parameter corresponding to this column. This choice would entail $\mathrm{RR}_1 = 1$ and $\mathrm{RR}^*_{2,2} = 1$, i.e., a multiplicative response model with allele 1 as the reference allele. If the model should allow deviations from the multiplicative response, including $1_M \cdot 1_F$ in the $\boldsymbol{X}$ matrix would provide an estimate of $\mathrm{RR}^*_{2,2}$. By similar constructions, all models described in this paper are covered. The exact form of the $\boldsymbol{X}$ matrix is not important for the likelihood derivation below.

**The asymptotic variance-covariance matrix** In our likelihood model, we write $l_N(\boldsymbol{\beta}) = \log(L_N(\boldsymbol{\beta}))$ for the log-likelihood based on $N$ triads. Let $\hat{\boldsymbol{\beta}}_N$ be the corresponding maximum likelihood estimator of the $p \times 1$ parameter vector $\boldsymbol{\beta}$. As described above, the $\boldsymbol{\beta}$ parameter vector contains information about haplotype frequencies and relative risks; typically, $\beta_i = \log(\mathrm{RR}_i)$ for some component $i$ of the vector, where $\mathrm{RR}_i$ is the relative risk associated with haplotype $h_i$. We denote the asymptotic $p \times p$ variance-covariance matrix by $\boldsymbol{\Sigma}$. From standard likelihood theory,

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

as $N \to \infty$ [1, Chapter 10]. The matrix $\boldsymbol{\Sigma}$ is given as the inverse of the expected information matrix, $\mathcal{I}(\boldsymbol{\beta})$, with element $(i,j)$ defined as

$$-E\left\{ \frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j} \right\},$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function [2].

If $N$ is the total number of observed triads, then $N$ is Poisson distributed with expected value equal to the sum over all rows, i.e., $m_{\boldsymbol{\cdot}}$. Conditional on $N$, the number of triads corresponding to a row in $\boldsymbol{G}$ (if they were fully observed) follows a multinomial distribution with cell probabilities $\boldsymbol{p}$. Hence, the likelihood contribution from a single observed (possibly ambiguous) triad $j$ is $\boldsymbol{a}_j^T\boldsymbol{p}$, and the full likelihood, accounting for ambiguities, is

$$L(\boldsymbol{\beta}) \propto m_{\boldsymbol{\cdot}}{}^N e^{-m_{\boldsymbol{\cdot}}} \prod_{j=1}^{N} \boldsymbol{a}_j^T \boldsymbol{p}.$$

The corresponding log-likelihood function is then

$$l(\boldsymbol{\beta}) = \sum_j (\log(\boldsymbol{a}_j^T \boldsymbol{m})) - m_{\boldsymbol{\cdot}}.$$

Applying the rules for vector differentials [3], we have that

$$\partial l(\boldsymbol{\beta}) = (\sum_j \boldsymbol{b}_j^T - \boldsymbol{m}^T)\boldsymbol{X}\partial\boldsymbol{\beta},$$

where

$$\boldsymbol{b}_j = \frac{\text{diag}(\boldsymbol{a}_j)\boldsymbol{m}}{\boldsymbol{a}_j^T \boldsymbol{m}}.$$

Furthermore, the second derivative of the log-likelihood function is

$$\partial^2 l(\boldsymbol{\beta}) = (\partial\boldsymbol{\beta})^T \boldsymbol{X}^T \left( \text{diag}(\sum_j \boldsymbol{b}_j) - \sum_j \boldsymbol{b}_j \boldsymbol{b}_j^T - \text{diag}(\boldsymbol{m}) \right) \boldsymbol{X}(\partial\boldsymbol{\beta}).$$

Consequently, the observed Fisher information matrix is

$$I_N(\boldsymbol{\beta}) = \boldsymbol{X}^T \left( \text{diag}(\boldsymbol{m}) - \text{diag}(\sum_j \boldsymbol{b}_j) + \sum_j \boldsymbol{b}_j \boldsymbol{b}_j^T \right) \boldsymbol{X},$$

and as $N \to \infty$,

$$\frac{1}{N} I_N(\boldsymbol{\beta}) \sim \boldsymbol{X}^T \left( \text{diag}(\boldsymbol{p}) - \text{diag}(E(\boldsymbol{b})) + E(\boldsymbol{b}\boldsymbol{b}^{\boldsymbol{T}}) \right) \boldsymbol{X}.$$

It follows that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_N$ is

$$\boldsymbol{\Sigma} = \left[\boldsymbol{X}^T \left(\mathrm{diag}(\boldsymbol{p}) - \mathrm{diag}(E(\boldsymbol{b})) + E(\boldsymbol{b}\boldsymbol{b}^{\boldsymbol{T}})\right) \boldsymbol{X}\right]^{-1},$$

and thus

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_N) \sim \frac{1}{N}\boldsymbol{\Sigma}.$$

## References

[1] Christensen R. Log-linear models and logistic regression. 2nd ed. New York, NY: Springer; 1997.

[2] Pawitan Y. In all likelihood. Oxford: Clarendon Press; 2001.

[3] Wand MP. Vector differential calculus in statistics. The American Statistician. 2002;56(1):55–62.

# Additional file 2 — Power and sample size calculations in Haplin

Haplin includes a complete setup for power calculations, extending beyond the single-SNP analyses of child, PoO and maternal effects. Here, we provide an extensive tutorial and illustrate power analyses for a range of relevant genetic and etiologic scenarios. Nevertheless, this tutorial is not intended as an exhaustive documentation of the power framework and its functions. We therefore strongly recommend consulting the **R** help page, which includes detailed and up-to-date information on the power functions and all their arguments. As we continue to expand our framework for power analysis, changes to the presented commands may occur. Such updates will be documented on the Haplin website at `https://people.uib.no/gjessing/genetics/software/haplin`, as well as on the **R** help page.

The supplementary material is structured as follows. We start with an introduction of the asymptotics-based functions `snpPower` and `snpSampleSize`, before continuing with an extended tutorial of `hapPowerAsymp`. The two last sections are devoted to `hapRun` and `hapPower`, which calculate the power by using simulations.

## `snpPower` and `snpSampleSize`

For single-SNP analyses of child effects, statistical power and sample size calculations are most easily done with the Haplin functions `snpPower` and `snpSampleSize`. `snpPower` computes the power for a single SNP by counting the number of "real" case alleles (transmitted alleles from case triads), "real" control alleles (all alleles from control triads) and pseudo-control alleles (non-transmitted alleles from case families). A multiplicative dose-response relationship is assumed. `snpPower` calculates the power by using the asymptotic normal approximation for the natural logarithm of the odds ratio (the relative risks and odds ratios are used interchangeably due to the "rare disease assumption"). It computes the power for a given number of case families, control families, relative risks, minor allele frequencies (MAFs) and type I error rates. For example, to compute the power for 200 case-parent triads and 100 control children, assuming a relative risk of 1.4, a minor allele frequency of 0.2, and a nominal significance level of 5%, use the command

```
snpPower(cases=list(mfc=200), controls=list(c=100), RR=1.4, MAF=0.2, alpha=0.05).
```

In `snpPower`, the power can be calculated for a mixture of different family designs and for several combinations of the input variables simultaneously. Please refer to the **R** help page and the Haplin website for an explanation of the arguments and its options.

Note that most of the functionality of `snpPower` is covered by the more flexible Haplin function `hapPowerAsymp`, which also extends to power analyses of haplotype effects, parent-of-origin (PoO) effects, maternal effects, gene-environment interactions (GxE), etc. However, `snpPower` is somewhat easier to apply and is therefore useful for simple power calculations of single-SNP child effects.

`snpSampleSize` is the inverse function of `snpPower`. For child effects, it computes the number of case and control families required for a single SNP to attain the desired power for specified family designs and given values of relative risks, minor allele frequencies and type I error rates. Examples and documentation are given on the **R** help page and on the Haplin website.

# hapPowerAsymp: Extensions to X-linked markers, gene-environment interactions and haplotype effects

Basic power calculations of child, PoO and maternal effects using `hapPowerAsymp` are described in the main article. Here we show how to extend power analyses to X-linked markers, GxE, and haplotype effects.

## X-linked markers

Genetic association analyses of X-linked markers might be of particular relevance if the prevalence of a complex trait or disease is systematically different for males versus females. Various X-chromosome models are implemented in Haplin. The models depend on the underlying assumptions regarding allele-effects in males versus females, which may include sex-specific baseline risks, shared or distinct relative risks for males and females, as well as X-inactivation in females. A detailed description of the parameterization models is provided in our previous studies [1, 2, 3]. Corresponding power analyses are readily available in `hapPowerAsymp`, and an example of X-chromosome power analysis is shown in Table S1a. In addition to the arguments needed to perform power calculations of child, maternal or PoO effects on autosomal markers, three arguments are required to specify an X-linked penetrance model. The argument `xchrom` must be set to `TRUE`, which indicates power analysis of X-chromosome markers. Furthermore, the argument `sim.comb.sex` specifies how to deal with sex differences on the X-chromosome. We have used the option `single`, which means that the effect of one (single) allele in males equals the effect of a single allele dose in females. However, the default value is `double`, which corresponds to X-inactivation; a single allele in males has the same effect as one of the two alleles in homozygous females, assuming that the other allele is inactivated. The argument `BR.girls` gives the ratio of baseline risk for females relative to males. In the example of Table S1a, we assumed a ratio of 1, i.e., the same baseline risk in females and males.

## Gene-environment interactions

A gene-environment interaction occurs when a genetic effect is modified by an environmental exposure. For example, maternal alcohol consumption, cigarette smoking or vitamin intake in the periconceptional period might modify the association between SNPs and a birth defect [4, 5]. The genetic effect in question might be a child, PoO or maternal effect. In Haplin, interactions between a genetic effect and a categorical exposure variable are incorporated into the log-linear framework by fitting the log-linear model separately for each exposure stratum. We then apply a Wald test to assess whether the relative risk estimates differ significantly across exposure levels [6, 7]. In `hapPowerAsymp`, the power to detect a GxE effect is automatically computed when the number of strata is larger than 1, specified by the argument `n.strata`. Each of the stratum-specific arguments `cases`, `controls`, `haplo.freq`, `RR`, `RRcm`, `RRcf` and `RR.mat` are given as lists. Their lengths should be equal to the number of strata, and each element of the list specifies the argument for one stratum. An example of GxE power analysis of child effects and two exposure strata is given in Table S1b. We used 500 case-parent triads in the first stratum and 300 case-parent triads in the second (`cases = list(c(mfc=500),c(mfc=300))`). The list format is, however, only needed for arguments that vary across strata. Here we assumed that the allele frequencies are the same in both strata, and the list format is therefore redundant (`haplo.freq = c(0.8,0.2)`). There are no associations in the first stratum, whereas the minor allele is associated with the disease in the second stratum (`RR = list(c(1,1),c(1,1.4))`).

## Haplotypes

By default, hapPowerAsymp performs power calculations for a diallelic SNP. However, the extension to haplotypes is straightforward but requires a basic understanding of how the haplotypes are generated in Haplin. The number of markers and haplotypes is determined by the vector nall, where the number of markers is equal to length(nall), and the number of different haplotypes is equal to prod(nall). Thus, two diallelic markers are denoted by nall = c(2,2), whereas a single marker with four alleles is denoted by nall = 4. The haplotypes are determined by creating all possible haplotypes from the given markers, in a sequence where the first marker varies most quickly. For instance, if nall = c(3,2), there are six haplotypes in total. Taken in order, the haplotypes are 1-1, 2-1, 3-1, 1-2, 2-2, and 3-2. If haplo.freq = c(0.3,0.05,0.1,0.1,0.2,0.25) and RR = c(1,2,1,1,1,1), haplotype 2-1 has a twofold risk compared to the rest of the haplotypes. Table S1c shows a haplotype example with two diallelic markers. Compared with the reference (by default the most frequent haplotype), all haplotypes are associated with an increased risk of disease. The power to detect the effect of an individual haplotype is calculated by analyzing that specific haplotype separately against the reference, using the Wald test with one degree of freedom. Here, the individual power estimates range between 63% and 74%. We also calculate the overall power, i.e., the power to detect any difference among the haplotypes, by analyzing the haplotypes jointly. With a total of four haplotypes, the Wald test has three degrees of freedom, and the power is approximately 84%.

The power calculations can be extended to three or more markers at a locus in a similar manner. An example of three diallelic SNPs (eight haplotypes) is provided in Table S1d.

## Other effects

The power analyses in Table S1 were calculated for child effects. However, the power to detect PoO effects is readily computed by replacing the relative risk argument RR by RRcm and RRcf, similar to the example in Table 3b of the main article. For instance, in the GxE example (Table S1b), replacing RR with RRcm=list(c(1,1),c(1,1)) and RRcf=list(c(1,1),c(1,2)) would mean that there is no risk associated with the allele transmitted from the mother in either stratum, whereas the paternally derived allele is associated with the disease only in the second stratum. Maternal effects are included by adding the argument RR.mat (see Table 3c and d of the main article for examples).

**Table S1** Asymptotic power calculations in Haplin

| Etiologic scenarios | Haplin commands | Output | |
|---|---|---|---|
| a) X-chromosome | hapPowerAsymp(cases = c(mfc=500),<br>haplo.freq = c(0.8,0.2), RR = c(1,1.2),<br>xchrom = T, sim.comb.sex = "single",<br>BR.girls = 1) | $haplo.power<br>Haplotype RR.power<br>1 ref<br>2 0.3 | |
| b) GxE | hapPowerAsymp(n.strata = 2,<br>cases = list(c(mfc=500),c(mfc=300)),<br>haplo.freq = c(0.8,0.2),<br>RR = list(c(1,1), c(1,1.4))) | $haplo.power<br>Haplotype RR.power<br>1 ref<br>2 0.47 | |
| c) Haplotype effects, two diallelic markers | hapPowerAsymp(nall=c(2,2), cases = c(mfc=500),<br>haplo.freq = c(0.4,0.3,0.2,0.1),<br>RR = c(1,1.3,1.4,1.5)) | $haplo.power<br>Haplotype RR.power<br>1-1 ref<br>2-1 0.63<br>1-2 0.74<br>2-2 0.73 | $overall.power<br>child<br>0.8370423 |
| d) Haplotype effects, three diallelic markers | hapPowerAsymp(nall=c(2,2,2), cases = c(mfc=500),<br>haplo.freq = c(0.25,0.2,0.1,0.1,0.05,0.1,0.1,0.1),<br>RR = c(1,1.2,1,1,1.2,1,1.7,1)) | $haplo.power<br>Haplotype RR.power<br>1 1-1-1 ref<br>2 2-1-1 0.24<br>3 1-2-1 0.05<br>4 2-2-1 0.05<br>5 1-1-2 0.12<br>6 2-1-2 0.05<br>7 1-2-2 0.89<br>8 2-2-2 0.05 | $overall.power<br>child<br>0.7416111 |

The power is calculated for a) X-chromosome models; b) GxE effects; c) Haplotype effects, two diallelic markers; d) Haplotype effects, three diallelic markers. The argument `cases` determines the study design and its sample size, and the argument `RR` specifies the relative risk associated with a child effect. Note that the order of alleles to which the relative risk parameters refer corresponds to the order used for the haplotype frequencies in argument **`haplo.freq`**. The most frequent allele/haplotype is used as reference. The arguments **`xchrom`**, **`sim.comb.sex`** and **`BR.girls`** are specific for power analyses of X-linked markers. The argument **`n.strata`** determines the number of strata and is required for power analyses of GxE effects. In c) and d), the number of markers and haplotypes is specified by the vector **`nall`**. Because the default value is 2 (corresponding to a diallelic SNP), this argument was not explicitly expressed in examples a) and b). The nominal significance level defaults to 0.05, but other values can be specified by the argument **`alpha`**

## Introduction to hapRun and hapPower

The function hapRun simulates genotype data and performs the subsequent statistical inference. The results can then be fed to hapPower, which calculates the actual power. Because hapRun performs both the simulations and the subsequent statistical inference, the aimed target effects must be specified in addition to the simulation-specific parameters. The commands therefore require knowledge of the functions haplin and haplinStrat, which perform the statistical inference within hapRun.

We demonstrate the power simulation functions hapRun and hapPower by using the same scenarios as in Table 3 of the main article, in addition to the GxE example in Table S1. The examples are shown in Table S2. Additional to the arguments provided in hapPowerAsymp, hapRun requires the arguments nall, RRstar and response to be specified (RRstar.mat must also be specified in order to simulate maternal effects). The vector nall specifies the number of markers and haplotypes. In hapPowerAsymp, nall has the default value 2, whereas in hapRun the argument must be given explicitly. Moreover, hapRun handles deviations from the multiplicative dose-response relationship. Such deviations can be simulated by the arguments RRstar and RRstar.mat, which correspond to the parameters $\mathrm{RR}^*$ and $\mathrm{RR}^{(M)*}$ in Eq (1-4) from Table 2 in the main article. The target effect is specified by the haplin argument response, which has the option "mult" for estimating a multiplicative dose-response relationship, and the option "free" for estimating separate single-dose and double-dose effects. In Table S2, we *simulate* and *test* a multiplicative dose-response relationship throughout (RRstar = c(1,1), RRstar.mat = c(1,1) and response = "mult"). However, if one were to forget response = "mult", the simulated data (following a multiplicative dose-response relationship), when fed to haplin, would be used to estimate separate single-dose and double-dose effects, corresponding to the haplin default value "free".

In hapRun, the arguments RRcm and RRcf must be specified in order to *simulate* PoO effects. However, to *test* for PoO effects, one also needs to specify poo = TRUE in hapRun (Table S2b). It is thus possible to simulate PoO effects without actually testing them. The same is true for maternal effects (Table S2c and d); the arguments RR.mat and RRstar.mat enable simulations of maternal effects, but this effect is not tested unless maternal = TRUE in hapRun.

The argument hapfunc specifies which Haplin function to run on the simulated data in hapRun. Because most genetic association analyses are conducted using the function haplin, hapfunc = "haplin" is the default value. However, GxE effects are analyzed using haplinStrat, as shown in Table S2e. We recommend consulting the **R** help files for a thorough description of these Haplin functions and for further information on the target effects and the arguments to be passed onto haplin and haplinStrat.

Right now the output of hapPower contains more information than the output of hapPowerAsymp. The first result column, overall.power, displays the power for detecting an overall difference between the null model (no effects) and the full model. Whereas the other results of hapPower are based on the Wald test, the overall result is based on the likelihood ratio test. RRdd.power and RRmdd.power show the power to detect a double-dose child effect or a double-dose maternal effect, respectively. Because we have assumed a multiplicative dose-response relationship, the power to detect a double-dose child or maternal effect equals the power to detect a single-dose effect. The multiplicative double-dose PoO effect is interpreted analogously to the multiplicative double-dose child effect but is estimated by stratifying on parental origin.

## **Table S2** Simulated power in Haplin

| Effects | Haplin commands | Output |
|---|---|---|
| a) Child | `res.child <- hapRun(nall = 2, cases = c(mfc=500),`<br>`        haplo.freq = c(0.8,0.2),`<br>`        RR = c(1,1.4), RRstar = c(1,1),`<br>`        response = "mult")`<br><br>`hapPower(res.child)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RR.power RRdd.power`<br>`       1         0.878      ref        ref`<br>`       2         0.878    0.877      0.877` |
| b) PoO | `res.PoO <- hapRun(nall = 2, cases = c(mfc=500),`<br>`        haplo.freq = c(0.8,0.2),`<br>`        RRcm = c(1,2), RRcf = c(1,1.5),`<br>`        RRstar = c(1,1), poo = T,`<br>`        response = "mult")`<br><br>`hapPower(res.PoO)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RRcm.power RRcf.power`<br>`       1         0.999        ref        ref`<br>`       2         0.999      0.999      0.839`<br><br>`RRcm_RRcf.power RRdd.power`<br>`            ref        ref`<br>`          0.482      0.999` |
| c) Child and maternal | `res.childmat <- hapRun(nall = 2, cases = c(mfc=500),`<br>`        haplo.freq = c(0.8,0.2), RR = c(1,1.4),`<br>`        RRstar = c(1,1), RR.mat = c(1,1.2),`<br>`        RRstar.mat = c(1,1), maternal = T,`<br>`        response = "mult")`<br><br>`hapPower(res.childmat)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RR.power RRdd.power`<br>`       1         0.921      ref        ref`<br>`       2         0.921    0.899      0.899`<br><br>`RRm.power RRmdd.power`<br>`      ref         ref`<br>`    0.419       0.419` |
| d) PoO and maternal | `res.PoOmat <- hapRun(nall = 2, cases = c(mfc=500),`<br>`        haplo.freq = c(0.8,0.2), RRcm = c(1,2),`<br>`        RRcf = c(1,1.5), RRstar = c(1,1),`<br>`        RR.mat = c(1,1.2), RRstar.mat = c(1,1),`<br>`        maternal = T, poo = T, response = "mult")`<br><br>`hapPower(res.PoOmat)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RRcm.power RRcf.power`<br>`       1             1        ref        ref`<br>`       2             1      0.995      0.637`<br><br>`RRcm_RRcf.power RRdd.power RRm.power RRmdd.power`<br>`            ref        ref       ref         ref`<br>`          0.209      0.999     0.164       0.164` |
| e) GxE | `res.GxE <- hapRun(nall = 2, n.strata = 2,`<br>`        cases = list(c(mfc=500),c(mfc=300)),`<br>`        haplo.freq = c(0.8,0.2),`<br>`        RR = list(c(1,1), c(1,1.4)),`<br>`        RRstar = c(1,1), response = "mult",`<br>`        hapfunc = "haplinStrat")`<br><br>`hapPower(res.GxE)` | `The power was calculated using 1000 of 1000 files`<br>`child`<br>`0.484` |

We simulate the power for a diallelic SNP, using a MAF of 0.2 (`haplo.freq = c(0.8,0.2)`). The study design and the corresponding sample size are determined by the argument `cases`. The arguments `RR` and `RRstar` specify the relative risks associated with the child effect. Note that a multiplicative dose-response model is simulated by `RRstar = c(1,1)` and tested by the argument `response = "mult"`. The power to detect a PoO effect is calculated by replacing `RR` by the two relative risk arguments `RRcm` and `RRcf`, which refer to the parental origin of the allele carried by the child. In addition, we need to include the target effect by the argument `poo = T`. A maternal effect can be simulated by adding the maternal relative risk parameters `RR.mat` and `RRstar.mat` to the original child or PoO command, and the effect is tested by adding `maternal = T`. Power calculations of GxE effects are conducted by adding the arguments `n.strata` and `hapfunc = "haplinStrat"`, as well as specifying the stratum-specific arguments. Note that the order of alleles to which the relative risk parameters refer corresponds to the order used for the haplotype frequencies. Here, the less frequent allele is set as the risk allele, and the more frequent allele is used as reference. The nominal significance level defaults to 0.05, but different levels can be specified by the argument `alpha`. By default, `hapRun` simulates 1000 replicates of data files. Other values can be set by the argument `n.sim`. As `hapRun` simulates genotype data, the results will vary, and the precision depends on the number of replicates. The simulation procedure is time-consuming. However, one can speed up the calculations by using parallel processing, specified by the argument `cpus`

## Extended family designs and missing individuals

The simulation procedure handles a variety of child-parent configurations. As shown in Table S3a, designs such as case-parent triads and case-mothers dyads may be combined. The argument `controls` can be extended in a similar manner. Moreover, if genotype data are missing at random, e.g., due to failed genotyping, missing case or control individuals can be generated at random through the arguments `gen.missing.cases` and `gen.missing.controls`. If the arguments are single numbers between 0 and 1, missing data are generated at random with these proportions for all case and controls individuals. In Table S3b, 10% of all case individuals (mothers, fathers and children) are missing. If the arguments are vectors of length equal to the number of markers, missing data are generated at random with the corresponding proportions for each marker. The arguments can also be matrices with the number of rows equal to the number of markers and three columns. Each row corresponds to a single marker, and the columns correspond to mothers, fathers and children, respectively. Thus, `gen.missing.cases = matrix(c(0,0.2,0),nrow=1)` simulates haplotype data in which 20% of the case fathers are missing at random. To ensure that the data are simulated correctly, it might be worthwhile to look at the simulated files. The argument `dire = "sim"` saves the simulated files to the given directory. If there is a large number of simulated files, a test run should be performed with a small number of data replicates (specified by the argument `n.sim`). We also note that the function `hapSim` can be used to simulate genotype data in Haplin format, without performing the actual testing.

**Table S3** Simulated power in Haplin- extended family designs and missing individuals

| | Haplin commands | Output |
|---|---|---|
| a) Extended family designs | `res.study.design <- hapRun(nall = 2,`<br>`        cases = c(mfc=400, mc=100),`<br>`        haplo.freq = c(0.8,0.2),`<br>`        RR = c(1,1.4), RRstar = c(1,1),`<br>`        response = "mult")`<br><br>`hapPower(res.study.design)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RR.power RRdd.power`<br>`       1         0.879      ref        ref`<br>`       2         0.879    0.878      0.878` |
| b) Missing individuals | `res.gen.missing <- hapRun(nall = 2,`<br>`        cases = c(mfc=500),`<br>`        gen.missing.cases = 0.1,`<br>`        haplo.freq = c(0.8,0.2),`<br>`        RR = c(1,1.4), RRstar = c(1,1),`<br>`        response = "mult")`<br><br>`hapPower(res.gen.missing)` | `The power was calculated using 1000 of 1000 files`<br>`  haplos overall.power RR.power RRdd.power`<br>`       1         0.826      ref        ref`<br>`       2         0.826    0.825      0.825` |

The power is simulated for a diallelic SNP, using different child-parent configurations and a MAF of 0.2 (`haplo.freq = c(0.8,0.2)`). In a), we combine 400 case-parent triads and 100 case-mother dyads to create one dataset (`cases = c(mfc=400, mc=100)`). In b), we simulate genotype data in which 10% of all case individuals are missing at random (`gen.missing.cases = 0.1`). The arguments `RR` and `RRstar` specify the relative risks associated with a child effect. A multiplicative dose-response model is simulated by `RRstar = c(1,1)` and tested by the argument `response = "mult"`. Note that the order of alleles to which the relative risk parameters refer corresponds to the order used for the haplotype frequencies. Here, the less frequent allele is set as the risk allele, and the more frequent allele is used as reference. The nominal significance level defaults to 0.05, but different levels can be specified by the argument `alpha`. By default, `hapRun` simulates 1000 replicates of data files. Other values can be set by the argument `n.sim`. As hapRun simulates genotype data, the results will vary, and the precision depends on the number of replicates. The simulation procedure is time-consuming. However, one can speed up the calculations by using parallel processing, specified by the argument `cpus`

# References

[1] Jugessur A, Skare Ø, Lie RT, Wilcox AJ, Christensen K, Christiansen L, et al. X-linked genes and risk of orofacial clefts: evidence from two population-based studies in Scandinavia. PLoS One. 2012;7(6):e39240.

[2] Skare Ø, Gjessing HK, Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, et al. A new approach to chromosome-wide analysis of X-linked markers identifies new associations in Asian and European case-parent triads of orofacial clefts. PLoS One. 2017;12(9):e0183772.

[3] Skare Ø, Lie RT, Haaland ØA, Gjerdevik M, Romanowska J, Gjessing HK, et al. Analysis of parent-of-origin effects on the X chromosome in Asian and European orofacial cleft triads identifies associations with *DMD*, *FGF13*, *EGFL6*, and additional loci at Xp22.2. Front Genet. 2018;9:25.

[4] Haaland ØA, Jugessur A, Gjerdevik M, Romanowska J, Shi M, Beaty TH, et al. Genome-wide analysis of parent-of-origin interaction effects with environmental exposure (PoOxE): an application to European and Asian cleft palate trios. PLoS One. 2017;12(9):e0184358.

[5] Haaland ØA, Lie RT, Romanowska J, Gjerdevik M, Gjessing HK, Jugessur A. A genome-wide search for gene-environment effects in isolated cleft lip with or without cleft palate triads points to an interaction between maternal periconceptional vitamin use and variants in *ESRRG*. Front Genet. 2018;9:60.

[6] Skare Ø, Jugessur A, Lie RT, Wilcox AJ, Murray JC, Lunde A, et al. Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. Ann Hum Genet. 2012;76(3):221–236.

[7] Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. Ann Hum Genet. 2018;82(2):60–73.

# Paper III

## Design efficiency in genetic association studies

Received: 25 April 2019 | Revised: 20 December 2019 | Accepted: 21 December 2019

DOI: 10.1002/sim.8476

RESEARCH ARTICLE

Statistics in Medicine WILEY

# Design efficiency in genetic association studies

**Miriam Gjerdevik**[1,2] | **Håkon K. Gjessing**[1,3] | **Julia Romanowska**[1,3] | **Øystein A. Haaland**[1] | **Astanand Jugessur**[1,2,3] | **Nikolai O. Czajkowski**[4,5] | **Rolv T. Lie**[1,3]

[1]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[2]Department of Genetics and Bioinformatics, Norwegian Institute of Public Health, Oslo, Norway

[3]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

[4]Department of Psychology, University of Oslo, Oslo, Norway

[5]Division of Mental Health, Norwegian Institute of Public Health, Oslo, Norway

**Correspondence**
Miriam Gjerdevik, Department of Global Public Health and Primary Care, University of Bergen, N-5020 Bergen, Norway.
Email: miriam.gjerdevik@uib.no

**Funding information**
Bergen Medical Research Foundation, 807191; The Research Council of Norway, 245464/F50, 262700

**Abstract**

Selecting the best design for genetic association studies requires careful deliberation; different study designs can be used to scan for different genetic effects, and each design has its own set of strengths and limitations. A variety of family and unrelated control configurations are amenable to genetic association analyses, including the case-control design, case-parent triads, and case-parent triads in combination with unrelated controls or control-parent triads. Ultimately, the goal is to choose the design that achieves the highest statistical power using the lowest cost. For given parameter values and genotyped individuals, designs can be compared directly by computing the power. However, a more informative and general design comparison can be achieved by studying the relative efficiency, defined as the ratio of variances of two different parameter estimators, corresponding to two separate designs. Using log-linear modeling, we derive the relative efficiency from the asymptotic variance of the parameter estimators and relate it to the concept of Pitman efficiency. The relative efficiency takes into account the fact that different designs impose different costs relative to the number of genotyped individuals. We show that while optimal efficiency for analyses of regular autosomal effects is achieved using the standard case-control design, the case-parent triad design without unrelated controls is efficient when searching for parent-of-origin effects. Due to the potential loss of efficiency, maternal genes should generally not be adjusted for in an initial genome-wide association study scan of offspring genes but instead checked post hoc. The relative efficiency calculations are implemented in our R package Haplin.

KEYWORDS

case-parent triad, Haplin, parent-of-origin effects, power and sample size, relative (Pitman) efficiency

## 1 | INTRODUCTION

Optimizing the design of a genetic association study requires careful consideration because (among other things) there are several factors to assess (eg, recruitment costs, genotyping costs, phenotypic costs, statistical power, and design-induced biases). The most common design for genetic association analysis is the standard case-control design in which individuals with and without the disease in question are genotyped. By contrast, if case-parent triad data are collected by genotyping cases and their biological parents, parent-of-origin (PoO) effects or direct effects of the maternal genome during fetal development (ie, maternal effects) can also be investigated.[1-4] Case-parent triads can also be combined with unrelated control-parent triads in a hybrid design.[5-9] Although the case-parent triad design is mostly used when the outcome occurs early in life, this design can be used for any condition, provided that parents are available for genotyping.

The statistical power is an important aspect of design comparison. Frequently, study designs are compared directly through a power analysis without considering the total number of individuals that needs to be genotyped. For example, a fixed number of complete case-parent triads could be compared with the same number of case-mother dyads. However, this approach ignores the costs of data collection. In this article, our objective is to present comparisons that enable the highest statistical power to be achieved using the smallest sample collection and assay costs. We assess this through the quantity known as relative efficiency, defined as the ratio of variances of estimators for the same parameter computed from two different designs, or equivalently, the ratio of the sample sizes needed for each of the two designs to achieve the same significance level and power. We demonstrate how the relative efficiency measures relate to the concept of Pitman efficiency.[10]

We have previously developed an extensive framework for genetic epidemiological analyses of binary traits based on log-linear modeling, implemented in the R package Haplin.[4,11-13] Haplin includes a complete setup for power and sample size calculation,[14,15] which is useful in study planning and in interpreting findings from a genome-wide association study (GWAS). In this article, we present a structured overview of different genetic effects and etiologic scenarios that are applicable to diseases with onset throughout the lifespan, along with appropriate choices of study designs. Our primary focus is on estimating the relative efficiency, which is readily assessed within the power calculation framework of Haplin.

The article is structured as follows. First, we introduce the relevant genetic effects and the family-based designs that are the focus of this article. Second, we describe our sampling and penetrance models, explain the concept of relative efficiency, and illustrate its association with statistical power. Finally, we study the relative efficiency of different designs for different genetic effects, both for single-nucleotide polymorphisms (SNPs) and for haplotypes, that is, the combinations of alleles from several SNPs within a locus. Although we focus on autosomal markers, the methodology presented is readily applicable to SNPs or haplotypes on the X chromosome. A discussion of relative efficiency is provided in Appendix A. In Appendix B, we provide a heuristic derivation of the relative efficiency for regular autosomal effects. To facilitate analysis of other genetic mechanisms, study designs, and input parameters, we provide Haplin commands for various scenarios on the Haplin website at `https://people.uib.no/gjessing/genetics/software/haplin`.

## 2 | BACKGROUND

The R package Haplin is a comprehensive framework for genetic association analyses of binary traits based on log-linear modeling.[4] It implements a full maximum-likelihood model for estimation and calculates explicit estimates of relative risks with asymptotic standard errors (SEs) and confidence intervals. Haplin enables the estimation of regular autosomal effects, PoO effects, and maternal effects, as well as interactions between genetic effects and categorical or ordinal exposure variables.[11,13] It allows for parallel processing of analyses as well as data structure for handling GWAS data. In Haplin, the main unit of analysis is the case-parent triad. However, the log-linear model can readily incorporate unrelated controls or control triads that are population-based (ie, of unknown disease status), or, under the rare disease assumption, unaffected controls or control triads.[7,16,17] Note that unrelated controls are optional since “pseudocontrols” in principle can be derived from the nontransmitted parental alleles in case-parent triads.[18-21] To account for unknown parent of origin in ambiguous (uninformative) triads, for example, when the mother, father, and child are all heterozygous for the same two alleles, Haplin uses the expectation maximization (EM) algorithm.[22] The EM algorithm also accounts for individuals that are missing “by design,” such as when case-parent triads are reduced to case-mother dyads due to missing data on fathers, assuming that the missingness is random, that is, independent of genotype. The log-linear model in Haplin assumes Mendelian transmission, Hardy-Weinberg equilibrium (HWE), and random mating, although moderate deviations from HWE are unlikely to cause bias.[23] A detailed description of the underlying model is provided in several of our previous publications.[4,11,13] For applications of Haplin to GWAS data, readers are referred to some of our previous publications.[24-29]

**TABLE 1** Overview of genetic effects available in Haplin

| Effects | Description |
|---|---|
| Regular autosomal | A regular autosomal effect is a standard effect of the offspring's own genes. It occurs when a variant allele inherited from one or both parents increases or decreases the risk of a condition. |
| PoO | A PoO effect occurs if the effect of a variant allele in an individual depends on whether it is inherited from the mother or from the father. Hypothetically, an allele might be protective when inherited from the mother but detrimental when inherited from the father. In statistical terms, we define a PoO effect as an interaction since the effect of an allele is modified by its parent of origin. In contrast, analyses of regular autosomal effects assume that the effect of an allele in an individual is independent of whether it is transmitted from the mother or the father. Note that genomic imprinting may cause PoO effects.[42,59] Imprinting is an epigenetic phenomenon where one of the inherited parental alleles is expressed whereas the other is silenced. |
| Maternal | A maternal genetic effect occurs when a variant allele carried by the mother increases or decreases the risk of a phenotype in her child, regardless of whether the allele has been inherited by the child or not.[34] It is expected to operate mainly via mechanisms in the intrauterine environment.[60] This is different from regular autosomal and PoO effects, where we estimate the effects of the child's own alleles. The relevance of maternal effects was recently demonstrated for an individual's educational attainment,[61] but may be particularly relevant for conditions that depend directly on fetal development. |

*Note*: Adapted from Gjerdevik et al.[14]
Abbreviation: PoO, parent-of-origin.

## 2.1 | Genetic effects

A GWAS scans the entire genome for common variants agnostically, without any prior information about the biological significance of a gene for the trait or disease under investigation. Hence, the selection of an appropriate design for a GWAS requires careful planning and depends heavily on the genetic effect being studied. Haplin enables the estimation of several genetic effects, and we focus here on regular autosomal, PoO, and maternal effects. Table 1 (adapted from Gjerdevik et al[14]) provides an explanation of the genetic effects.

## 2.2 | Study designs

### 2.2.1 | The case-control design

Similar to classic epidemiological studies of environmental and behavioral risk factors, the case-control design is often used in genetic association analyses (Figure 1A). The allele frequencies of cases and controls are contrasted to identify variants associated with the trait or disease, and familiar methods such as logistic regression and chi-squared tests can be used to discover associations.[30] The case-control design is efficient in uncovering regular autosomal effects and their interactions with exposure or stratification variables such as environmental risk factors, study sites, and ethnicity. However, population stratification might lead to spurious associations if not controlled for.

### 2.2.2 | The case-parent triad and dyad designs

The case-parent triad design involves genotyped cases and their biological parents and is based on the observation that parental genotypes of affected offspring could be used to study associations between a disease and allelic variants.[31,32] For regular autosomal effects, the frequencies of alleles transmitted to cases are compared to the frequencies of

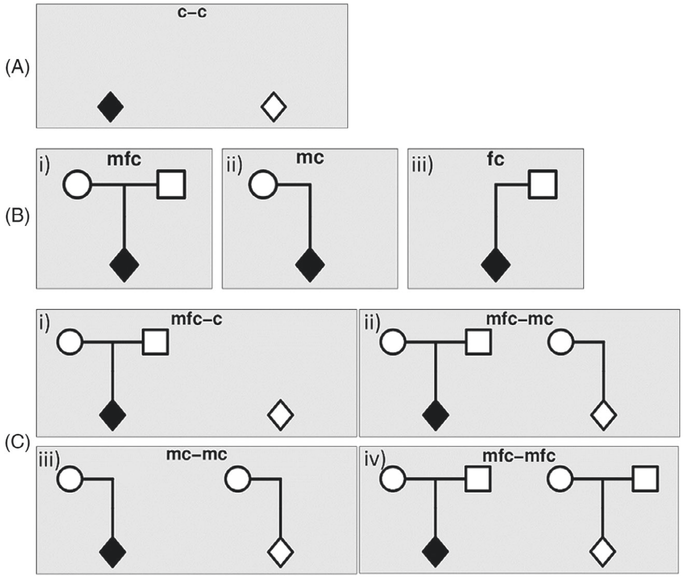

**FIGURE 1** A selection of designs for genetic association analyses: (A) case-control design (c-c); (B) various case-parent designs: (i) case-parent triad (mfc); (ii) case-mother dyad (mc); (iii) case-father dyad (fc); (C) a selection of hybrid designs: (i) case-parent triad with independent control (mfc-c); (ii) case-parent triad with independent control-mother dyad (mfc-mc); (iii) case-mother dyad with independent control-mother dyad (mc-mc); (iv) case-parent triad with independent control-parent triad (mfc-mfc)

non-transmitted (pseudocontrol) alleles. Hence, the case-parent triad design does not rely on independent controls and is protected against population stratification since the relevant information is extracted from within-family contrasts. Since Spielman et al[33] proposed the transmission disequilibrium test (TDT) for genetic association testing, exploring family-based designs and their utility for studying different types of genetic effects has been an intense area of research for several decades. Truncated versions of the case-parent triad design have been introduced, with the case-mother and case-father dyad designs comprising genotyped cases and their biological mothers or fathers, respectively. The various constellations are illustrated in Figure 1B. With information on parental genotypes, the case-parent triad and dyad designs allow the estimation and testing of PoO or maternal effects. For PoO and maternal effects, Connolly and Heron[34] reviewed different statistical methodologies and compared them according to statistical power and their suitability for studying different etiologic scenarios. Methods for testing PoO effects include extensions of the TDT approach, such as the transmission-asymmetry test (TAT) and the parental-asymmetry test (PAT),[3] conditional logistic regression,[20,21] and log-linear[1-4] and multinomial modeling.[17,35,36] With the exception of TAT and PAT, these approaches can also account for maternal effects.[34] Despite the inherent strengths of the case-parent triad and dyad designs, there are also some drawbacks. One such drawback is that they rely on Mendelian transmission. Another limitation is that, without independent controls, it is impossible to estimate the main effect of an environmental exposure. There might also be practical concerns, such as obtaining DNA from parents if the disease in question is late onset.

### 2.2.3 | The hybrid design

To combine the advantages of the case-control and the family-based designs, joint analyses of various combinations of case-parent triads and unrelated controls in a hybrid design have been proposed.[5-7,17,37] An overview of hybrid designs has been provided by Infante-Rivard et al,[38] and different configurations are illustrated in Figure 1C. The full hybrid design comprises complete pairs of case-parent triads and control-parent triads, but truncated versions may include case-parent triads supplemented by control-mother dyads[9] or case-mother dyads supplemented by control-mother dyads.[8,39] Analysis methods such as log-linear and multinomial modeling approaches are particularly appealing as they can readily be adapted to accommodate the broad spectrum of various hybrid designs as well as a wide array of causal scenarios and genetic effects.[11,13,17,24,27,29,35,40] As an example, they can easily be extended to include the maternal-fetal genotype incompatibility test.[41] Nevertheless, although the hybrid design combines the merits of both the case-control and case-parent designs, a straightforward combined analysis may still be influenced by population stratification or non-Mendelian transmission.

### 2.2.4 | Notation

We use the abbreviations provided in Figure 1 to describe the study designs. The letters c, m, and f denote the child (case or control), mother, and father, respectively. The left side of the hyphen denotes case families, whereas the right side denotes control families. For instance, mfc denotes the case-parent triad, whereas mfc-c denotes a hybrid design consisting of case-parent triads and unrelated controls (ie, the control parents have not been genotyped). We will use the term hybrid design to describe all constellations of study designs consisting of case families and independent control families, except for the straightforward c-c design. Although a case together with a control dyad or control triad can be seen as a hybrid design, these designs are rare in practice and will not be discussed.

# 3 | METHODS

## 3.1 | Parameterization of penetrances

We have developed a complete setup for power and sample size calculations in Haplin.[14] The calculations can be performed analytically using the asymptotic variance-covariance structure of the parameter estimator or by a straightforward simulation procedure. Relative efficiency is easily assessed within this framework, and the basic calculations are for regular autosomal, PoO, and maternal effects, with the results depending on the underlying parameterization models. The penetrance models, that is, the probability of a child having the disease conditional on a specific genetic composition, are defined in Table 2 (adapted from Gjerdevik et al[14]). For regular autosomal effects, the penetrance model is parameterized as $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl}$, where $B$ serves as a baseline parameter, and $\mathrm{RR}_j$ is the relative risk associated with allele $A_j$. The double-dose parameter $\mathrm{RR}^*_{jl}$ measures the deviation from what would be expected in a multiplicative dose-response relationship, that is, $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and $\mathrm{RR}^*_{jl} = 1$ when $j \neq l$. The double-dose estimates provide information about the effect of allele dose on risk. For a diallelic SNP with reference allele $A_1$, the penetrance model can written as $P(D|A_1A_1) = B$, $P(D|A_1A_2) = B \cdot \mathrm{RR}$ and $P(D|A_2A_2) = B \cdot \mathrm{RR}^2\mathrm{RR}^* = B \cdot \tilde{\mathrm{R}}\mathrm{R}$. A recessive effect of $A_2$ would then be seen as $\mathrm{RR} = 1$ and $\tilde{\mathrm{R}}\mathrm{R} \neq 1$, a dominant effect would mean that $\mathrm{RR} = \tilde{\mathrm{R}}\mathrm{R} \neq 1$, and a multiplicative dose-response relationship would be seen as $\tilde{\mathrm{R}}\mathrm{R} = \mathrm{RR}^2$ (see Gjessing and Lie).[4]

Since a mother and her child have one allele in common, maternal effects might be statistically confounded with regular autosomal or PoO effects of the child's own genes.[42,43] An important feature of the log-linear model is, therefore, the possibility of incorporating and adjusting for maternal effects. Specifically, maternal effects can be addressed simultaneously with regular autosomal or PoO effects by including the maternal risk parameters, as outlined in Table 2. Statistically, we are thus able to separate the effects of maternal alleles from the effect of maternally-derived alleles carried by the offspring.

**TABLE 2** Parameterization of penetrances

| Effects | Parameterization of Penetrances |
|---|---|
| Regular autosomal | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl}$ |
| PoO | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl}$ |
| Regular autosomal and maternal | $B \cdot \mathrm{RR}_j \mathrm{RR}_l \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ |
| PoO and maternal | $B \cdot \mathrm{RR}_{M,j} \mathrm{RR}_{F,l} \mathrm{RR}^*_{jl} \cdot \mathrm{RR}^{(M)}_i \mathrm{RR}^{(M)}_j \mathrm{RR}^{(M)*}_{ij}$ |

*Note*: $B$ is the baseline risk level associated with the (more frequent) reference allele.
$\mathrm{RR}_j$ is the risk increase or decrease associated with allele $A_j$, relative to $B$.
$\mathrm{RR}_{M,j}$ and $\mathrm{RR}_{F,j}$ are the relative risks associated with allele $A_j$, depending on whether the allele is derived from the mother or the father, respectively. Here, we define a PoO effect as the relative risk ratio $\mathrm{RRR}_j = \mathrm{RR}_{M,j}/\mathrm{RR}_{F,j}$, which is a measure of the risk increase (or decrease) associated with $A_j$ when the allele is transmitted from the mother as opposed to from the father.
$\mathrm{RR}^*_{jl}$ estimates deviations from the risk that would be expected in a multiplicative dose-response relationship, that is, $\mathrm{RR}^*_{jl} = \mathrm{RR}^*_j$ when $j = l$ and $\mathrm{RR}^*_{jl} = 1$ when $j \neq l$.
$\mathrm{RR}^{(M)}_i$ is the relative risk associated with allele $A_i$ carried by the mother, and $\mathrm{RR}^{(M)*}_{ij}$ is the maternal double-dose parameter, with an interpretation analogous to $\mathrm{RR}^*_{ij}$.
We set RR = 1 for the reference allele to ensure that the model is not overparameterized.
Adapted from Gjerdevik et al.[14]
Abbreviation: PoO, parent-of-origin.

We consider a multiplicative dose-response relationship throughout this article, that is, $\text{RR}_j^*$ is kept fixed at 1 for all $j$ and only $\text{RR}_j$ is estimated (an analogous interpretation applies for the parameterizations of PoO and maternal effects in Table 2). The estimation of $\text{RR}_j^*$ is possible and would allow other response models, for example, recessive or dominant, but these situations are not explored herein.

The statistical inference of the log-linear model in Haplin is based on log-transformed relative risks and relative risk ratios using the Wald test. We calculate the relative efficiency based on the asymptotic variance-covariance structure of the parameter estimator, and a derivation of the asymptotic variance-covariance matrix is given by Gjerdevik et al.[14] However, the simulation procedure in Haplin is equally applicable and has been shown to provide similar results within the range of sample sizes and allele frequencies usually studied.[14] For external validation, the power calculation modules in Haplin have previously been compared with the power attained in data simulations by EMIM (Estimation of Maternal, Imprinting, and interaction effects using Multinomial modelling),[17,35,36] which is another well-established tool for the estimation of various genetic effects based on genotype data from a number of different child-parent configurations. The consistency observed between Haplin and EMIM for regular autosomal, PoO, and maternal effects demonstrates the computational accuracy of the inference methods used in both programs and suggests that power and relative efficiency calculations in Haplin are applicable to genetic association studies based on either log-linear or multinomial modeling.[14]

## 3.2 | Asymptotic relative efficiency

Power analysis allows for a comparison of different designs when all parameter values have been specified. It demonstrates the possible scope of a study, that is, what is feasible logistically, and should, therefore, be an essential part of study planning. However, for "global" comparisons of statistical tests, relative efficiency is a more useful measure. In statistical terms, the relative efficiency of two designs is defined as the ratio of sample sizes required for each of the designs to attain the same significance level and power. This is equivalent to the ratio of variances of two different parameter estimators, corresponding to two separate study designs, taking into account that different designs require a different number of individuals to be genotyped. Figure 2 illustrates the relationship of relative efficiency to sample size and power. For regular autosomal effects, the efficiency of the c-c design is approximately 1.5 relative to the mfc design, which is well known from other studies.[44] For instance, if 1200 individuals (600 cases and 600 controls) are needed to reach a power of 0.8 with the c-c design, 1800 individuals (600 case-parent triads) are required with the mfc design to achieve the same power.

For the purpose of this article, we aim to compare tests asymptotically. Consider the problem of testing the null hypothesis $H_0 : \beta = 0$ versus the alternative $H_1 : \beta \neq 0$ for a fixed nominal level, $\alpha$, where $\beta$ is the log relative risk. With a given sample size $N$, the power of the test converges to 1 as $|\beta| \to \infty$. Similarly, when $\beta$ is fixed, the power converges to 1 as $N \to \infty$. The limiting power functions are identical for all reasonable tests, and such an approach is, therefore, unhelpful. When $N$ increases, the minimum detectable effect size decreases. To make an informative comparison of different designs,

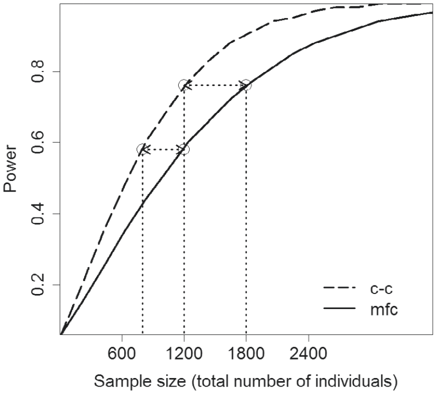

**FIGURE 2** Relative efficiency derived from power and sample size. Here, we compare the efficiency of the c-c design relative to the mfc design for regular autosomal effects. The power is calculated for a diallelic SNP at the 5% nominal significance level, using a MAF of 0.2 and an RR of 1.3. The sample size $N$ is defined as the total number of individuals, that is, $N = 1800$ means either 900 cases and 900 controls or 600 case-parent triads. If $N = 1200$, the power is nearly 0.8 for the c-c design. However, approximately $N = 1800$ individuals are required for the mfc design to reach the same power. Similarly, we need $N = 800$ individuals for the c-c design to attain an approximate power of 0.6, whereas $N = 1200$ individuals are required for the mfc design. Hence, the efficiency of the c-c design is 1.5 compared with the mfc design

we, therefore, examine the power at alternatives that approach the null hypothesis, that is, we shrink the alternative as $N$ increases, making it harder to discriminate between the null and alternative hypotheses as the number of observations increases. This is known as the Pitman efficiency,[10] and an explanation of this concept is provided in Appendix A. Most effect sizes reported from genetic association studies of complex traits are small, and empirical studies show that individual relative risks of disease are commonly below two.[45-48] Intuitively, the Pitman efficiency is thus a reasonable measure of the asymptotic relative efficiency in our setting.

### 3.3 | Analyses

We define $k$ : 1 as the ratio of control families to case families, regardless of the number of individuals within each family. If $k = 0.5$, we have twice as many case families as control families. For example, for the mfc-mc design, we might have 100 control-mother dyads and 200 case-parent triads. Our main results pertain to the relative efficiency, and we present it here as a function of $k$ on the log-scale. The efficiencies of various study designs are compared with that of the case-parent triad design (mfc), that is, we use the case-parent triad design as a "reference design." As mentioned previously, the relative efficiency will take into account the total number of genotyped individuals within each design. For example, 150 case-mother dyads are compared with 100 case-parent triads. If $k = 1$, a hybrid design with 50 case-parent triads and 50 control-parent triads is compared with 100 case-parent triads, and if $k = 2$, a hybrid design with 50 case-parent triads and 100 control-parent triads is compared with 150 case-parent triads. Only the ratio of control families to case families, not the actual number of control and case families, affects the relative efficiency estimates.

In genetic association studies, it makes sense to integrate data collection and assay costs with the concept of relative efficiency. For example, if the recruitment of case children occurs at a hospital where parents are likely to be present, parental pseudocontrols would be less expensive than independent controls. However, when studies are nested within a cohort that has already been sampled, the costs of genotyping DNA samples are typically considered equal for all individuals. Hence, for the majority of this article, the data collection costs are simply defined as the number of genotyped individuals. That is, we assume the same costs for all individuals, independent of the individual being a child, mother or father, case or control. However, differential costs of data collection may occur if, for instance, publicly available reference samples (eg, from catalogs such as the Wellcome Trust Case Control Consortium,[49] the UK Biobank,[50] and the Norwegian Mother, Father and Child Cohort Study[51,52]) are included in the study. As a special scenario, we analyze situations in which controls or control families are available without additional costs. For all analyses, we consider well-defined and clinically verified phenotypes, thus ignoring the costs of phenotyping.

The analyses were performed using the Haplin relative efficiency calculator `hapRelEff`. The results were obtained under the null hypothesis, corresponding to the Pitman efficiency.[10] However, we note that relative efficiency estimates in Haplin can also be obtained under alternative (nonnull) hypotheses, and investigators can readily apply our functions to study how alternative effect estimates relevant to their own research question would affect the relative efficiency values.

## 4 | RESULTS

### 4.1 | Regular autosomal effects

Figure 3 illustrates the relative efficiency for regular autosomal effects as a function of $k$, using two different values of the minor allele frequency (MAF). We used the mfc design as the reference, to which the other designs were compared. Unless the ratio of controls to cases is highly skewed, we see that the c-c design provides the best results. The optimal relative efficiency is achieved when $k = 1$. Moreover, we observe that the mfc design is more efficient than the mc or fc design. This result is independent of $k$, as no control families are sampled. Note that the contribution of a case mother or control mother is equal to the contribution of a case father or control father, respectively. We also see that the relative efficiencies of the hybrid designs decrease when two or three individuals are included in the control family. This is also observed when $k$ becomes sufficiently large. Furthermore, for designs consisting of case dyads or control dyads, that is, mc, fc, mc-mc, fc-fc, mfc-mc, and mfc-fc, the relative efficiency is influenced by the MAF. The MAF does not affect the relative efficiency of the c-c, mfc-c, and mfc-mfc designs.

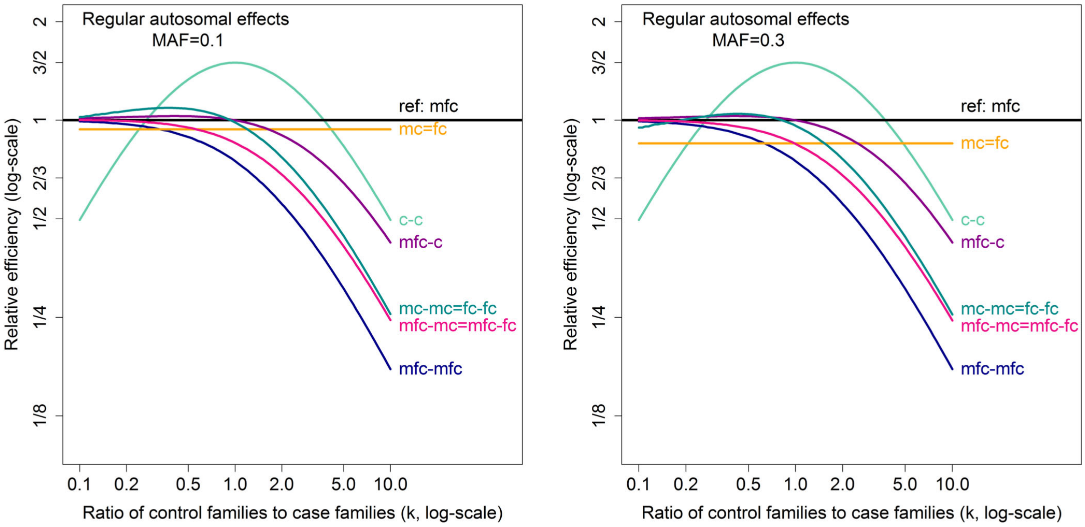

**FIGURE 3** Relative efficiency of regular autosomal effects for a given ratio of control families to case families ($k$). The efficiency of different study designs is compared with that of the case-parent triad design (mfc) under the null hypothesis of RR=1. The equality sign (eg, mc=fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

A heuristic formula for the relative efficiency of regular autosomal effects is derived in Appendix B. Equation (B1) verifies the results of Figure 3, and an inspection of the formula provides a better understanding of the observed relationships between the different study designs and each genotyped individual.

## 4.2 | PoO effects

Figure 4 shows the relative efficiency for PoO effects as a function of $k$. Again, we compared the relevant study designs with the mfc design under the null hypothesis of $\text{RRR} = \text{RR}_M = \text{RR}_F = 1$. When the MAF is 0.1 (left panel), the mc and fc designs are more efficient than the mfc design. However, this relationship reverses when the MAF is 0.3 (right panel). PoO effects are primarily estimated in case families, by comparing the frequency of alleles transmitted from mother to child with the frequency of alleles transmitted from father to child. Hence, the relative efficiency decreases when $k$ increases or when the number of genotyped individuals within a control family increases. Moreover, the relative efficiencies of the mfc-c, mfc-mc, mfc-fc, and mfc-mfc designs are not influenced by the MAF.

## 4.3 | Maternal effects

A putative maternal effect detected in a genome-wide scan may, at closer inspection, turn out to be caused by alleles carried by the offspring.[42,43] In Haplin, maternal effects are therefore assessed while accounting for the effects of the offspring's own alleles (see Table 2). Figure 5 shows the relative efficiency for maternal effects as a function of $k$ while adjusting for possible regular autosomal effects (left panel) and PoO effects (right panel). The results were calculated under the global null, that is, all relative risks are equal to one, using a MAF of 0.1. Overall, the mfc design is a good choice when adjusting for regular autosomal effects. However, when adjusting for PoO effects, a hybrid design generally performs better for small values of $k$. In both panels, the relative efficiency of the hybrid designs decreases when the number of genotyped individuals within a control family increases, as well as when $k$ becomes sufficiently large. This was also seen in the above analyses of regular autosomal and PoO effects. Note that we excluded the mc, fc, and fc-fc designs when adjusting for PoO effects because the models based on these designs would become overparameterized.

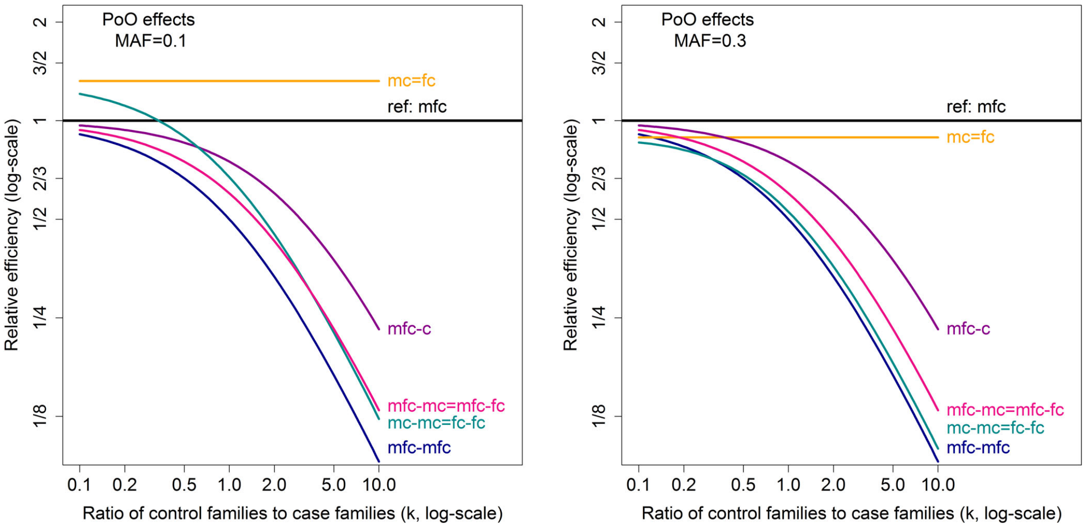

**FIGURE 4** Relative efficiency of PoO effects for a given ratio of control families to case families ($k$). The efficiency of different study designs is compared with that of the case-parent triad design (mfc) under the null hypothesis of RRR=$\mathrm{RR}_M$=$\mathrm{RR}_F$=1. The equality sign (eg, mc=fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

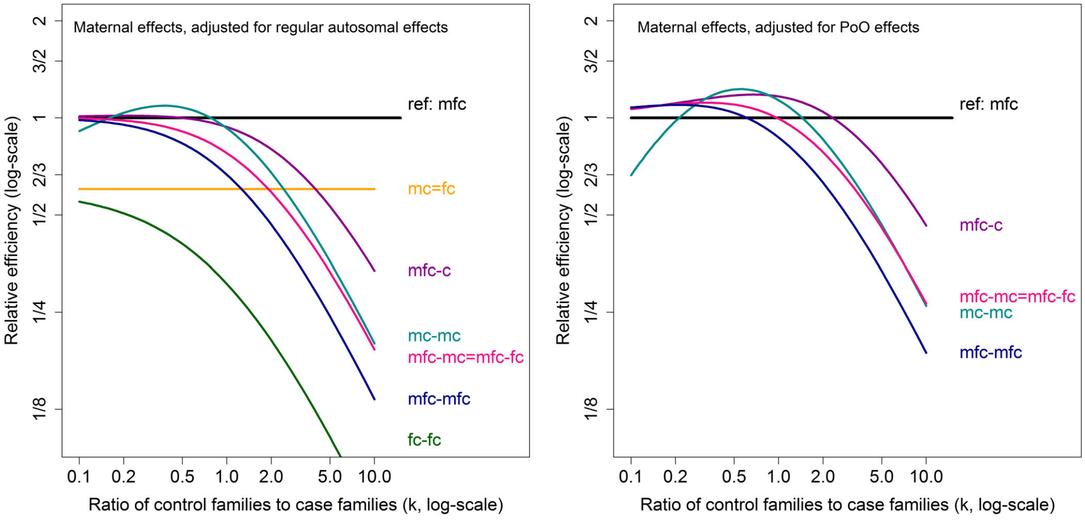

**FIGURE 5** Relative efficiency of maternal effects for a given ratio of control families to case families ($k$). The efficiency of different study designs is compared with that of the case-parent triad design (mfc) under the global null (ie, all RRs are equal to 1). We assumed a MAF of 0.1. The equality sign (eg, mfc-mc=mfc-fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

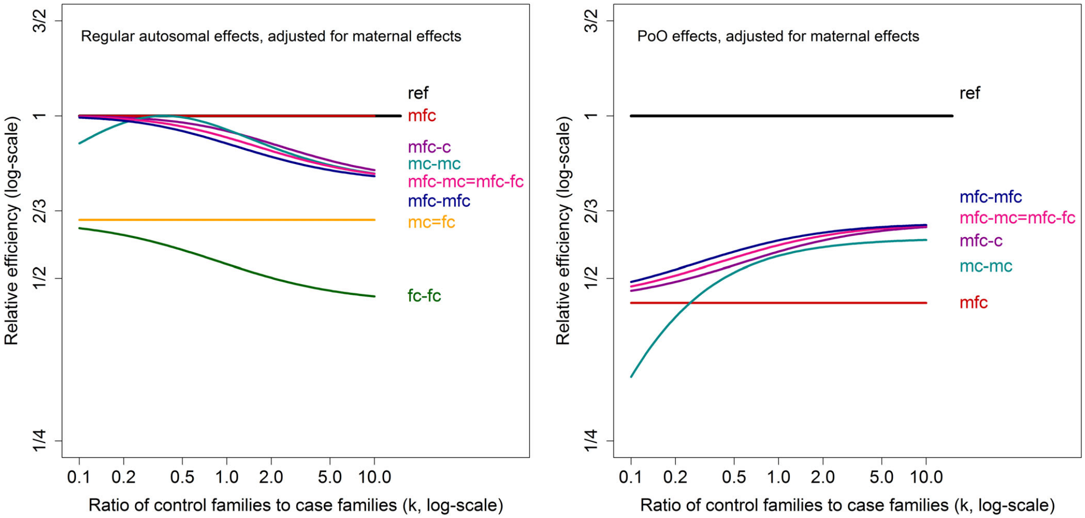

**FIGURE 6** Relative efficiency when adjusting for maternal effects for a given ratio of control families to case families ($k$). For each design, we first adjusted for maternal effects under the global null (ie, all RRs are equal to 1). We then repeated the analysis without adjusting for maternal effects and compared the results. The unadjusted analyses were used as references. We assumed a MAF of 0.1. The equality sign (eg, mfc-mc=mfc-fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

## 4.4 | Relative efficiency when adjusting for maternal effects

Including a search for maternal effects in a full GWAS analysis is likely to reduce the power to detect regular autosomal or PoO effects. Figure 6 demonstrates this loss of efficiency as a function of $k$ for regular autosomal effects (left panel) and PoO effects (right panel). We used a MAF of 0.1 in both panels. For each design, we first adjusted for possible maternal effects (even though we did not assume maternal effects in the parameterization model in Table 2, ie, we set $RR^{(M)} = 1$). We then repeated the analysis without adjusting for maternal effects and compared the results. The unadjusted analyses were used as references, and the mfc design is thus no longer a global reference. For regular autosomal effects, adjusting for maternal effects generally decreases the efficiency. However, no loss in efficiency is observed for the mfc design. Although the genotypes of individuals and their mothers are correlated in the population, their contributions to the mfc analysis are close to orthogonal.[1,2] That is, the estimation of maternal parameters does not affect the estimation of regular autosomal parameters or their SEs, and little bias is introduced for the mfc design (results not shown). When searching for PoO effects, adjusting for maternal effects causes a substantial loss of power for all designs. The efficiency is more than halved for the mfc design.

## 4.5 | Haplotype reconstruction

The fundamental model in Haplin relates to a single multiallelic locus but extends directly to haplotypes, that is, the sequence of alleles from several closely linked markers within a locus, by statistically reconstructing unknown haplotype phase using the EM algorithm.[4] A haplotype analysis should enhance the possibility of enclosing a causal variant if the haplotype has a SNP on each side of the variant. However, this analysis might lose power due to haplotype reconstruction and an increased number of degrees of freedom.

In order to assess the relative efficiency when haplotype reconstruction is performed, we considered a situation where one marker with four alleles was compared with two diallelic SNPs. In both scenarios, there were four possible haplotypes (alleles 1, 2, 3, and 4 and SNP-haplotypes 1-1, 2-1, 1-2, and 2-2), with haplotype frequencies 0.1, 0.3, 0.3,

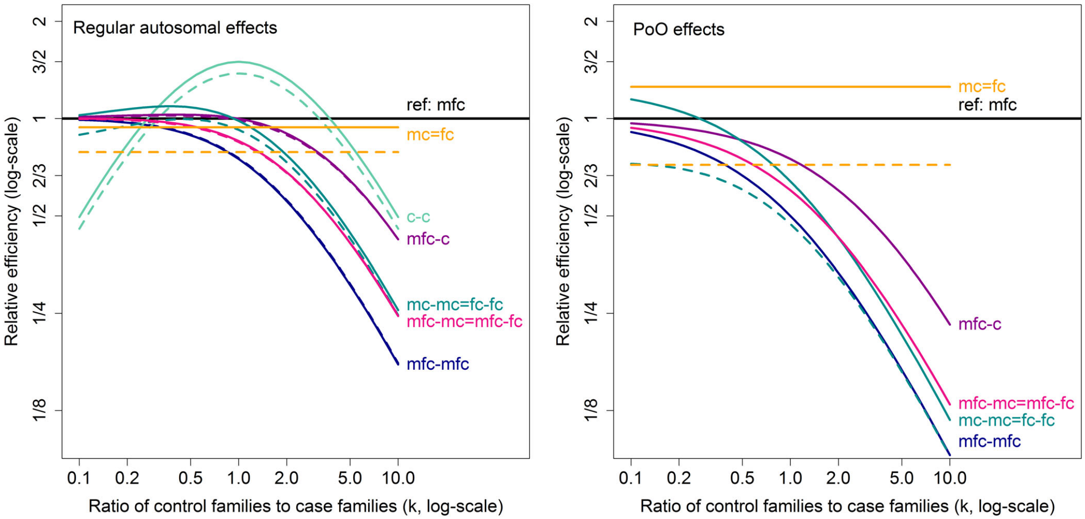

**FIGURE 7** Relative efficiency when haplotype reconstruction is performed for a given ratio of control families to case families ($k$). We constructed four alleles (haplotypes) from a single marker (alleles 1, 2, 3, and 4), and four haplotypes from two diallelic SNPs (haplotypes 1-1, 2-1, 1-2, and 2-2), both with haplotype frequencies 0.1, 0.3, 0.3, and 0.3, respectively, under the global null. A comparison of the solid (single marker, known phase) and dashed (haplotypes from two diallelic SNPs) lines demonstrates the loss of efficiency for the least frequent haplotypes due to haplotype reconstruction, relative to the mfc design. Allele 4 and haplotype 2-2 were used as references. The equality sign (eg, mc=fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

and 0.3, respectively. The alleles are directly observed when derived from a single multiallelic marker, and a haplotype reconstruction is only needed in the analysis of haplotypes from multiple markers. In Figure 7, we considered the efficiency of the least frequent haplotype in all designs, relative to the mfc design, and assessed both regular autosomal and PoO effects. Allele 4 and haplotype 2-2 were chosen as references, respectively. As phase is unknown, haplotype reconstruction for the c-c design is purely a statistical reconstruction. However, if the data from an individual and one or both parents are available at a single locus, the parent of origin can be deduced directly unless all individuals are heterozygous for the same two alleles, such that the EM algorithm is only needed for these ambiguous dyads or triads. Designs that include case-parent triads are, therefore, less vulnerable to unknown phase than the c-c, mc, fc, mc-mc, and fc-fc designs. These findings are in general agreement with those of Douglas et al[53] and Schaid.[54] Note that, in general, the results depend on the haplotype frequencies and also on the reference haplotype (results not shown). The haplotype frequencies used in the example deviate little from their values under linkage equilibrium ($r^2 = 0.0625$). Thus, our analysis demonstrates a larger loss of efficiency than what would be expected when the SNPs are in close linkage disequilibrium. Moreover, haplotype reconstruction in Haplin depends partly on the HWE assumption. Deviations from this assumption can be assessed within the Haplin framework, but such investigations are beyond the scope of this article.

## 4.6 | The use of external control samples

It has become increasingly common to utilize data from external and publicly available reference or control samples.[49-52] Figure 8 illustrates the gains in relative efficiency when external controls or control families are added to the mfc design. The efficiency of the different hybrid designs is compared with that of the mfc design, and the controls are here considered to be free of cost. For regular autosomal effects, we see that the use of freely available control samples increases the efficiency. For PoO effects, however, it has been shown elsewhere that unrelated control samples would not increase the power attained by the mfc design alone.[14] Thus, the relative efficiency of the mfc-c, mfc-mc, mfc-fc, and mfc-mfc designs is equal to 1 for all values of $k$.

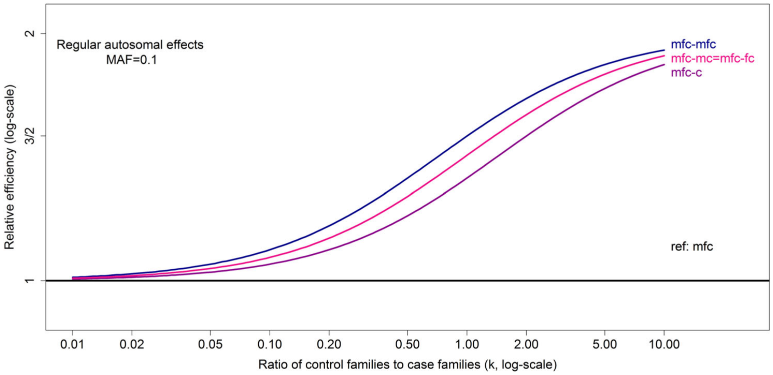

**FIGURE 8** Relative efficiency of regular autosomal effects for a given ratio of control families to case families ($k$). The efficiency of different hybrid designs is compared with that of the case-parent triad design (mfc) under the null hypothesis of RR=1. We consider the control samples to be free of charge, that is, without any sampling or genotyping costs. The equality sign (mfc-mc=mfc-fc) denotes that the two designs are interchangeable in terms of relative efficiency [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE 3** Application of relative efficiency to cleft palate data

| Effects | SNP | MAF[a] | Variance | | Empirical Relative Efficiency[b] | Theoretical Relative Efficiency[b] |
|---|---|---|---|---|---|---|
| | | | Case-mother dyads | Case-parent triads | | |
| Regular autosomal | rs2274616 | 0.1 | 0.0387 | 0.0370 | 0.96 | 0.94 |
| | rs12119556 | 0.3 | 0.0184 | 0.0154 | 0.84 | 0.85 |
| PoO | rs12137004 | 0.1 | 0.0649 | 0.0868 | 1.34 | 1.32 |
| | rs2357649 | 0.3 | 0.0399 | 0.0364 | 0.91 | 0.89 |

[a] Approximate estimates.

[b] The case-parent triad design is used as reference.

Abbreviations: MAF, minor allele frequency, PoO, parent-of-origin; SNP, single-nucleotide polymorphism.

## 4.7 | Application of Haplin to cleft palate only data

Cleft palate only (CPO) is a common craniofacial birth defect in humans, typically classified as to whether the cases occur with (nonisolated) or without (isolated) other congenital anomalies or identifiable malformation syndromes. The overall prevalence of isolated CPO is 5.0 per 10 000 births.[55] From our previously published GWAS,[56,57] genotype data from 550 isolated CPO families were available, including 466 complete case-parent triads. These families were primarily of European and Asian ancestry, although other ethnicities were also present in the data. The GWAS data set is available at the dbGaP database (https://www.ncbi.nlm.nih.gov/gap) under accession ID phs000094.v1.p1, and information on quality control and detailed characterizations of study participants have been provided elsewhere.[25] Background information on the study is given in the original publication,[56] and ethics approvals were obtained from the respective ethics committees for all the data in the cleft consortium.

To illustrate what the relative efficiencies may amount to with typical MAFs and effect sizes from our example data, we selected a total of 450 complete case-parent triads and chose SNPs with varying MAFs and effect estimates (RR or RRR) close to one for both regular autosomal and PoO effects. For case-mother dyads, the fathers were simply set to missing. To ensure an equal number of genotyped individuals for each design, 300 case-parent triads were randomly drawn from the 450 families using bootstrapping with 101 repetitions. The empirical relative efficiency was then calculated by dividing the median variance of the 101 case-parent triad replicates by the variance of the case-mother dyads. The results are displayed in Table 3, and the findings are in general agreement with the asymptotic calculations shown in Figures 3 and 4.

# 5 | ADDITIONAL CONSIDERATIONS

## 5.1 | Gene-environment interactions

A gene-environment interaction (GxE) occurs when a genetic effect is modified by an environmental exposure or a stratification factor such as ethnicity. For example, maternal exposures such as alcohol consumption, smoking, or vitamin intake during the periconceptional period might modify the association between SNPs and a birth defect.[25,28,29] Interactions between genetic effects and categorical exposure variables are incorporated into the log-linear framework of Haplin by fitting the log-linear model separately for each exposure stratum. A Wald test is then applied to detect whether the relative risk estimates differ significantly across exposure levels.[11,13] The genetic effect in question might be a regular autosomal, PoO, or maternal effect. Thus, GxE effects can be estimated for all study designs but are restricted to the genetic effects enabled by that design. Note, however, that the main effects of an environmental exposure cannot be estimated from the case-parent triad or dyad design alone without the addition of independent controls.

Because the GxE test stratifies on exposure levels, detecting a GxE effect requires a larger sample size than detecting the genetic effect alone. The SE of a GxE effect is determined by the standard errors of the individual genetic effects in the unexposed and exposed strata.[13] Provided that the same study design and parameter values are used in each stratum, the relative efficiency estimates are, therefore, directly transferable to GxE effects. Calculated under the global null, that is, $\text{RRR} = \text{RR}_{\text{exposed}} = \text{RR}_{\text{unexposed}} = 1$, Figures 3-5 would also apply to the relative efficiency for GxE effects in these situations.

## 5.2 | X-chromosome analysis

Haplin allows for analyses of X-linked markers, with corresponding PoO, maternal, and GxE effects. Genetic association analyses of X-linked markers are especially relevant if the prevalence of a complex trait differs systematically between males and females. In Haplin, different X-chromosome models may be fitted depending on the underlying assumptions, including sex-specific baseline risks, shared or different relative risks for males and females, and X-inactivation in females.[24,40] The methodology presented herein on relative efficiency is readily transferable to genetic effects on X-linked markers. Nevertheless, a discussion regarding sex effects is needed. For instance, when searching for X-linked PoO effects, females are needed to be able to compare maternally- and paternally-derived X-chromosome alleles. However, male individuals and fathers contribute to estimating allele frequencies.[13,27] They also facilitate haplotype reconstruction because phase can be deduced directly from fathers.

# 6 | CONCLUDING REMARKS

Statistical power is often a limiting factor for genetic association studies, and no comprehensive software has been available for the full assessment of power and comparison of study designs in such analyses to date. In this article, we provided insights into how relevant designs compare in terms of relative efficiency for a wide range of genetic effects and etiologic scenarios. Furthermore, we illustrated the methodology with extensive analyses and presented results for regular autosomal, PoO, and maternal effects. To facilitate the analysis of power and relative efficiency, the calculations have been implemented in our R package Haplin.[15]

The results herein relate to power and efficiency considerations only. Using either a single-SNP or a haplotype approach, the c-c design is recommended when the aim is to search for regular autosomal effects. An equal number of cases and controls maximizes the efficiency. However, additional correction for population stratification may be necessary for the c-c design. For a PoO analysis, the mfc design would be an overall good choice. Note that unrelated control families would not improve the power obtained by the case-parent triad design, as PoO effects are primarily estimated in case families by comparing the frequencies of alleles transmitted from mother to child with the frequencies of alleles transmitted from father to child.[14] Nonetheless, inferences based on the case-parent triad design rely on key assumptions that cannot be fully checked or corrected for without the inclusion of unrelated control families. For maternal effects, the mfc design is appropriate when adjusting for regular autosomal effects, whereas the mfc-c or mc-mc design would be a good choice when adjusting for PoO effects.

Due to the potential loss of power, we do not generally recommend including maternal effects in a full GWAS investigation of regular autosomal or PoO effects. Instead, we suggest additional post-scan analyses to control for possible confounding from maternal effects. As a matter of routine, the most promising SNPs from a GWAS analysis should be further examined for maternal effects.[25] However, we note that complex but less likely scenarios where maternal effects cancel out regular autosomal or PoO effects may go undetected by this strategy.

When analyzing real data, one would typically use a combination of several study designs. For example, the data can consist of case-parent triads supplemented by unrelated cases and controls.[37,38] Such mixture designs are readily handled in Haplin, both in the analysis module and in the power simulation module, but were not illustrated in this article.

The relative efficiency depends on multiple factors, such as the genetic effect in question, the MAF of a given SNP, and the study design. The results are, therefore, hard to summarize. Moreover, the most efficient design to test one hypothesis (ie, casual scenario) is not necessarily the best for testing another hypothesis. If different hypotheses about the modes of inheritance are to be tested, one may prefer a design that is reasonably efficient for a majority of hypotheses rather than the optimal design for a single hypothesis. Hence, since the mfc design is reasonably efficient for the genetic effects studied herein, it may be considered an overall optimal design. The importance of sampling case-parent triads is further strengthened since unrelated, ethnically matched controls have become more easily accessible through publicly available reference samples.[49-52]

The concluding recommendations in this article are subject to the log-linear model with the given assumptions, the investigated parameter values, and study designs; they should, therefore, not be interpreted as universal guidelines. Furthermore, practical issues should always be considered, such as the availability of case-parents or suitable controls, as well as recruitment and phenotyping costs. Nevertheless, the methodology presented herein is a useful approach toward optimizing the statistical power using the lowest sample collection and assay cost, and a careful assessment of possible study designs should be routinely performed prior to conducting a GWAS.

## ACKNOWLEDGEMENTS

The authors acknowledge four anonymous reviewers for their valuable comments on the article. Support for this work was provided by the Bergen Medical Research Foundation (BMFS) (Grant 807191), and by the Research Council of Norway (RCN) through Biobank Norway (Grant 245464/F50), and the Centres of Excellence funding scheme (Grant 262700). The funding bodies played no role in the design of the study, analysis or interpretation of data, nor in writing the manuscript.

## CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

## AUTHOR CONTRIBUTIONS

M.G. developed the relative efficiency and power calculation tools in Haplin, conceived, planned, and performed the analyses and drafted the manuscript. J.R., Ø.A.H., A.J., N.O.C., and R.T.L. helped develop the concepts and revised the manuscript. J.R. has also contributed to the recent developments of Haplin. H.K.G. developed the Haplin software, conceived, and planned the analyses and revised the manuscript. All the authors read and approved the final manuscript.

## DATA ACCESSIBILITY

Haplin is implemented in the statistical software R and can be installed from the official R package archive, CRAN (`https://cran.r-project.org`).[58] Standard power calculations in Haplin can be carried out analytically using the asymptotic variance-covariance structure of the parameter estimator (recently implemented in the function `hapPowerAsymp`), or else by a straightforward simulation approach (see functions `hapRun` and `hapPower`). Relative efficiency estimates are readily computed using the function `hapRelEff`.[15] For a thorough description of the Haplin functions and their arguments, please refer to the website at `https://people.uib.no/gjessing/genetics/software/haplin`. The CPO GWAS data are available at the dbGaP database (https://www.ncbi.nlm.nih.gov/gap) under accession ID phs000094.v1.p1.

## ORCID

*Miriam Gjerdevik* https://orcid.org/0000-0002-2604-2132
*Julia Romanowska* https://orcid.org/0000-0001-6733-1953
*Øystein A. Haaland* https://orcid.org/0000-0001-5288-7879

## REFERENCES

1. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet*. 1998;62(4):969-978.
2. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol*. 1998;148(9):893-901.
3. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet*. 1999;65(1):229-235.
4. Gjessing HK, Lie RT. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Ann Hum Genet*. 2006;70(3):382-396.
5. Nagelkerke NJD, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet*. 2004;12(11):964-970.
6. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet*. 2005;76(4):592-608.
7. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet*. 2005;77(4):627-636.
8. Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. *Am J Epidemiol*. 2008;168(5):541-547.
9. Vermeulen SH, Shi M, Weinberg CR, Umbach DM. A hybrid design: case-parent triads supplemented by control-mother dyads. *Genet Epidemiol*. 2009;33(2):136-144.
10. Noether GE. On a theorem of Pitman. *Ann Math Stat*. 1955;26(1):64-68.
11. Skare Ø, Jugessur A, Lie RT, et al. Application of a novel hybrid study design to explore gene-environment interactions in orofacial clefts. *Ann Hum Genet*. 2012;76(3):221-236.
12. Jugessur A, Skare Ø, Harris JR, Lie RT, Gjessing HK. Using offspring-parent triads to study complex traits: a tutorial based on orofacial clefts. *Nor Epidemiol*. 2012;21(2):251-267.
13. Gjerdevik M, Haaland ØA, Romanowska J, Lie RT, Jugessur A, Gjessing HK. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. *Ann Hum Genet*. 2018;82(2):60-73.
14. Gjerdevik M, Jugessur A, Haaland ØA, et al. Haplin power analysis: a software module for power and sample size calculations in genetic association analyses of family triads and unrelated controls. *BMC Bioinform*. 2019;20(1):165.
15. Gjessing HK. Haplin: analyzing case-parent triad and/or case-control data with SNP haplotypes *R Package Version* 7.1.0, 2019.
16. Weinberg CR, Shi M. The genetics of preterm birth: using what we know to design better association studies. *Am J Epidemiol*. 2009;170(11):1373-1381.
17. Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol*. 2011;35(1):19-45.
18. Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet*. 1993;52(6):1085-1093.
19. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet*. 1993;53(5):1114-1126.
20. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol*. 2004;26(3):167-185.
21. Cordell HJ. Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genet Epidemiol*. 2004;26(3):186-205.
22. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B*. 1977;39(1): 1-38.
23. Wise AS, Shi M, Weinberg CR. Family-based multi-SNP X chromosome analysis using parent information. *Front Genet*. 2016;7:20.
24. Skare Ø, Gjessing HK, Gjerdevik M, et al. A new approach to chromosome-wide analysis of X-linked markers identifies new associations in Asian and European case-parent triads of orofacial clefts. *PLoS One*. 2017;12(9):e0183772.
25. Haaland ØA, Jugessur A, Gjerdevik M, et al. Genome-wide analysis of parent-of-origin interaction effects with environmental exposure (PoOxE): an application to European and Asian cleft palate trios. *PLoS One*. 2017;12(9):e0184358.
26. Moreno ULM, Fomina T, Munger RG, et al. A population-based study of effects of genetic loci on orofacial clefts. *J Dent Res*. 2017;96(11):1322-1329.
27. Skare Ø, Lie RT, Haaland ØA, et al. Analysis of parent-of-origin effects on the X chromosome in Asian and European orofacial cleft triads identifies associations with *DMD*, *FGF13, EGFL6, and additional loci at Xp22.2. Front Genet*. 2018;9:25.
28. Haaland ØA, Lie RT, Romanowska J, Gjerdevik M, Gjessing HK, Jugessur A. A genome-wide search for gene-environment effects in isolated cleft lip with or without cleft palate triads points to an interaction between maternal periconceptional vitamin use and variants in *ESRRG*. *Front Genet*. 2018;9:60.
29. Haaland ØA, Romanowska J, Gjerdevik M, Lie RT, Gjessing HK, Jugessur A. A genome-wide scan of cleft lip triads identifies parent-of-origin interaction effects between *ANK3* and maternal smoking, and between *ARHGEF10* and alcohol consumption [Version 2]. *F1000Res*. 2019;8(960).

30. Clayton D, Hills M. *Statistical Models in Epidemiology*. Oxford, UK: Oxford University Press; 1993.
31. Falk CT, Rubinstein P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet*. 1987;51(3):227-233.
32. Self SG, Longton G, Kopecky KJ, Liang K-Y. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics*. 1991;47(1):53-61.
33. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*. 1993;52(3):506-516.
34. Connolly S, Heron EA. Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Brief Bioinform*. 2015;16(3):429-448.
35. Howey R, Cordell HJ. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. *BMC Bioinform*. 2012;13:149.
36. Howey R, Mamasoula C, Töpf A, et al. Increased power for detection of parent-of-origin effects via the use of haplotype estimation. *Am J Hum Genet*. 2015;97(3):419-434.
37. Stewart WCL, Cerise J. Increasing the power of association studies with affected families, unrelated cases and controls. *Front Genet*. 2013;4:200.
38. Infante-Rivard C, Mirea L, Bull SB. Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am J Epidemiol*. 2009;170(5):657-664.
39. Wang S, Yu Z, Miller RL, Tang D, Perera FP. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Hum Hered*. 2011;71(3):196-208.
40. Jugessur A, Skare Ø, Lie RT, et al. X-linked genes and risk of orofacial clefts: evidence from two population-based studies in Scandinavia. *PLoS One*. 2012;7(6):e39240.
41. Sinsheimer JS, Palmer CGS, Woodward JA. Detecting genotype combinations that increase risk for disease: the maternal-fetal genotype incompatibility test. *Genet Epidemiol*. 2003;24(1):1-13.
42. Hager R, Cheverud JM, Wolf JB. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics*. 2008;178(3):1755-1762.
43. Buyske S. Maternal genotype effects can alias case genotype effects in case-control studies. *Eur J Hum Genet*. 2008;16(7):783-785.
44. Cordell HJ, Clayton DG. Genetic association studies. *Lancet*. 2005;366(9491):1121-1131.
45. Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *Am J Epidemiol*. 2006;164(7):609-614.
46. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet*. 2007;39(1):17-23.
47. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*. 2007;17(10):1520-1528.
48. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753.
49. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678.
50. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.
51. Magnus P, Birke C, Vejrup K, et al. Cohort profile update: the Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol*. 2016;45(2):382-388.
52. Helgeland Ø, Vaudel M, Juliusson PB, et al. Genome-wide association study reveals dynamic role of genetic variation in infant and early childhood growth. *Nat Commun*. 2019;10:4448.
53. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet*. 2001;28(4):361-364.
54. Schaid DJ. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet Epidemiol*. 2002;23(4):426-443.
55. Mossey PA, Castilla EE. *Global Registry and Database on Craniofacial Anomalies*. Geneva, IN: World Health Organization; 2003.
56. 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*. *Nat Genet*. 2010;42(6):525-529.
57. Shi M, Murray JC, Marazita ML, et al. Genome wide study of maternal and parent-of-origin effects on the etiology of orofacial clefts. *Am J Med Genet A*. 2012;158(4):784-794.
58. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019.
59. Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. *Nat Rev Genet*. 2013;14(9):609-617.
60. Guilmatre A, Sharp AJ. Parent of origin effects. *Clin Genet*. 2012;81(3):201-209.
61. Kong A, Thorleifsson G, Frigge ML, et al. The nature of nurture: effects of parental genotypes. *Science*. 2018;359(6374):424-428.
62. Lehmann EL. *Elements of Large-Sample Theory*. New York, NY: Springer; 1999.
63. van der Vaart AW. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press; 2000.

64. Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley; 2013.
65. Schaid DJ. Disease-marker association. In: Elston R, Olson J, Palmer L, eds. *Biostatistical Genetics and Genetic Epidemiology*. Wiley reference series in biostatistics. West Sussex, UK: Wiley; 2002:206-217.

**How to cite this article:** Gjerdevik M, Gjessing HK, Romanowska J, et al. Design efficiency in genetic association studies. *Statistics in Medicine*. 2020;1–19. https://doi.org/10.1002/sim.8476

## APPENDIX A

### A1 Relative efficiency

For comparisons of statistical tests, the asymptotic relative efficiency is a useful measure.[62] The asymptotic relative efficiency is defined as the ratio of the asymptotic variances of two different estimators of the same parameter. Under general conditions (see Theorem 14.19 of Reference 63), this ratio corresponds to the ratio of sample sizes needed to achieve the same precision from the two different estimators, or the ratio of sample sizes needed to achieve the same significance level and power for two hypothesis tests about the parameter. In our setting, we compare the variances of the estimators of the same parameter computed from two different study designs, weighted by the number of genotyped individuals within each design. The weights allow us to compare the relative efficiency of the two different designs, subject to the constraint that each design contains the same number of genotyped samples. The relative efficiency thus refers to a ratio of the number of genotyped *individuals*, not a ratio of the number of *families.* Let $n$ denote the number of family structures with a case child. As $n$ varies, we assume the composition of family structures remains the same, relatively speaking. That is, we assume, for instance, that the ratio of case-parent triads to control-mother dyads remains the same, likewise for the ratio of case-mother dyads to complete control-parent triads, and so on.

#### A1.1 The asymptotic SE of the log-scale parameter estimator

In Haplin, we use the Wald test to conduct post-hoc inference on the log-transformed relative risk parameters, based on asymptotic normality (see Chapter 1.3 of Reference 64). The main univariate outcome measure is the log relative risk of the relevant genetic effect, that is, $\beta = \log(\text{RR})$. For PoO effects, the parameter of interest is the ratio of two relative risks, which means looking at the difference between the corresponding $\beta$ values, so the theory is the same. Based on the standard maximum likelihood theory, Haplin computes the SE $\sigma_n(\hat{\beta})$ of $\hat{\beta}$ from the observed Fisher information, using all available data, that is, with $n$ cases. If the composition of family structures is kept fixed as $n$ increases, we have that $\sqrt{n}\sigma_n(\hat{\beta}) \approx \omega(\beta)$, where $\omega(\beta)$ is the asymptotic SE of $\hat{\beta}$ computed from the Fisher information in the maximum likelihood model.[14] The value of $\omega(\beta)$ can thus be seen to represent a sample with only one case ($n = 1$). For instance, in a setting with 200 case-parent triads and 100 control-parent triads, $\omega(\beta)$ would, theoretically, correspond to a family structure with one case triad and half a control triad. The derivation of the asymptotic multivariate variance-covariance matrix is provided in a previous article.[14]

#### A1.2 Asymptotic relative efficiency

The asymptotic SE is characteristic of the design used in the estimation. When comparing two designs 0 and 1, with design 0 as reference, the asymptotic relative efficiency of design 1 over design 0, that is, using design 0 as reference, is

$$\left\{\omega^{(0)}(\beta)/\omega^{(1)}(\beta)\right\}^2 \cdot \frac{m_0}{m_1}, \tag{A1}$$

where $m_0$ and $m_1$ are the number of individuals to be genotyped in designs 0 and 1, respectively. For instance, the asymptotic relative efficiency of the case-control design over the case-parent triad design uses $m_0 = 3$ and $m_1 = 2$ (the case-parent triad design is used as reference). Note that a ratio larger than one favors design 1.

Comparing the asymptotic variances of estimators of the same parameter from different designs provides an intuitive understanding of relative efficiency. Alternatively, one can consider relative efficiency in terms of hypothesis testing. Consider the problem of testing the null hypothesis $H_0 : \beta = 0$ versus the alternative $H_1 : \beta \neq 0$ for a fixed nominal level. Let $\gamma_n(\beta)$ be the power of the Wald test based on $n$ cases (ie, $n$ family structures with one case child in each). Clearly, for a fixed alternative $\beta \neq 0$, $\lim_{n\to\infty}\gamma_n(\beta) = 1$. That is, with enough data, a relative risk RR different from one will eventually be detected by increasing the sample size sufficiently. To make an informative asymptotic comparison of two tests, that is, of tests for the same null hypothesis but based on two different designs, it is better to compare the efficiency of the tests when testing steadily decreasing effect sizes as the sample size increases. Here, we let the alternative to be tested for be $\beta_n = h/\sqrt{n}$, where $h$ is a fixed constant. Under general conditions,

$$\lim_{n\to\infty} \gamma_n(\beta_n) = \gamma(h),$$

where $\gamma(h)$ is the so-called local limiting power function (see Theorem 14.7 of Reference 63). In our setting, the limiting power function $\gamma$ of the Wald test with level $\alpha$ can be written

$$\gamma(h) = 1 - F_{\lambda(h)}(\chi^2_\alpha),$$

where $\chi^2_\alpha$ is the upper-$\alpha$ quantile of the chi-squared distribution with one degree of freedom and $F_\lambda$ is the cumulative distribution function of a one degree of freedom non-central chi-squared distribution, with $\lambda = \lambda(h)$ as the noncentrality parameter. The noncentrality parameter can be expressed as $\lambda(h) = (h/\omega(0))^2$, where $\omega(0)$ is the asymptotic SE of $\hat{\beta}$ under the null hypothesis. Hence, comparing two parameter estimators corresponding to different study designs is equivalent to comparing the locally attained power of the Wald test. That is, the asymptotic relative efficiency of two designs when testing the null hypothesis can be found from Equation (A1) by setting $\beta = 0$. Note that Equation (A1) is independent of $\alpha$ and $h$ when $\beta = 0$ (see Theorem 14.19 of Reference 63). This type of asymptotic relative efficiency for hypothesis tests is referred to as the Pitman efficiency.[10]

# APPENDIX B

## B1 An explicit formula for the asymptotic relative efficiency of regular autosomal effects for a diallelic SNP under $H_0$

For regular autosomal analyses of a diallelic SNP under $H_0$, a formula for the relative efficiency is easily derived by heuristic arguments. We quantify the statistical contribution of a genotyped individual by its "design factors" and count the effective number of cases and controls while assuming a multiplicative dose-response relationship. The case (affected individual) forms the basis of the family-based designs and is always assumed to be genotyped. We, therefore, define the effective number of cases as $n_1 = 1$. The total effective number of controls can be written as $n_0 = d_1 + kd_0$, where $d_1$ is the effective number of controls from a case family, $d_0$ is the effective number of controls from a control family, and $k : 1$ is the ratio of control families to case families.

A single case or control (without their genotyped parents) identifies only two case or control alleles, respectively. Hence, the design factors are $d_1 = 0$ and $d_0 = 1$. However, a single case-parent triad encompasses four alleles, two of which are inherited by the case child, two of which are not. The nontransmitted parental alleles form the so-called pseudocontrols.[20,21] Effects are seen as a contrast between the alleles of the pseudocontrols and the cases, similar to the approach used with a regular case-control design. A case-parent triad thus represents one case and one control ($d_1 = 1$). Conversely, a complete control-parent triad adds a single control offspring. Moreover, a pseudocontrol can also be formed, effectively resulting in two controls ($d_0 = 2$). Because these two controls together carry the same alleles as their parents, there is no need to genotype the original control child when both control parents have been genotyped.[7]

The issue of determining the design factor gets more complex when case dyads or control dyads are genotyped. If the case and only one of his/her parents are available, there are two case alleles and one control allele. However, deciding which of the parent's two alleles should be the control allele is not always possible when the other parent is missing. This results in a loss of efficiency, which leads to a design factor $d_1 < 1/2$, depending on the minor allele frequency (MAF).[65] If only one of the control parents is available for genotyping, genotyping the control offspring and his/her parent produces three control alleles. However, similar to the case-dyad scenario, if both the control offspring and his/her parent are heterozygous, one cannot distinguish which allele has been transmitted from the genotyped parent. Again, this leads to a loss of efficiency and a design factor $d_0 < 3/2$. The results are summarized in Table B1.

### B1.1 An explicit formula

The total (actual) number of genotyped individuals is equal to $G = l_1 + kl_0$, where $l_1$ and $l_0$ are the number of genotyped individuals within a case and control family, respectively, with the possible values 0, 1, 2, or 3. Under $H_0$, the SE of the

**TABLE B1** Design factors for regular autosomal effects under $H_0$ in the single-SNP situation

| | Control Family | | Case Family | |
|---|---|---|---|---|
| | $d_0$ | $l_0$ | $d_1$ | $l_1$ |
| MFC | 2 | 3 | 1 | 3 |
| MF | 2 | 2 | — | — |
| MC or FC | (3 – MAF · (1 – MAF))/2 [a] | 2 | (1 – MAF · (1 – MAF))/2 [a] | 2 |
| M or F | 1 | 1 | — | — |
| C | 1 | 1 | 0 | 1 |

*Note*: $d_0$ is the effective number of controls from a control family; $l_0$ is the number of genotyped individuals within a control family; $d_1$ is the effective number of controls from a case family; $l_1$ is the number of genotyped individuals within a case family.

Abbreviations: MAF, minor allele frequency.

[a] The effective number of controls is derived by subtracting the subset of ambiguous dyads.

difference between cases and controls is expected to be proportional to $\sqrt{1/n_0 + 1/n_1}$. Because $n_1 = 1$, the effective sample size for design $i$ can be written as

$$N_i \propto \frac{1}{\text{SE}_i^2} \propto \frac{n_0}{n_0 + 1}.$$

Relative to the number of genotyped individuals, the effective sample size for design $i$ is

$$\frac{N_i}{G_i}(k) = \frac{d_1 + kd_0}{(d_1 + kd_0 + 1)(l_1 + kl_0)}.$$

In this article, the case-parent triad design (mfc) is used as the reference, and we have that $\frac{N_{\text{mfc}}}{G_{\text{mfc}}}(k) = \frac{1}{6}$. Under $H_0$, the efficiency of design $i$ relative to the mfc design is thus

$$\frac{N_i/G_i}{N_{\text{mfc}}/G_{\text{mfc}}}(k) = \frac{6(d_1 + kd_0)}{(d_1 + kd_0 + 1)(l_1 + kl_0)}. \tag{B1}$$

When $k = 1$, we see that the relative efficiency is 3/2 for the case-control (c-c) design and 3/4 for the full hybrid (mfc-mfc) design, independent of the MAF. This corresponds to the results of Figure 3 in this article.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

UNIVERSITAS BERGENSIS

uib.no

ISBN: 9788230843086 (print)
9788230846315 (PDF)