Using offspring-parent triads to study complex traits: A tutorial based on orofacial clefts

Astanand Jugessur[1,2], Øivind Skare[1,3], Jennifer Ruth Harris[1], Rolv Terje Lie[3,1] and Håkon Kristian Gjessing[1,3]

1) Division of Epidemiology, Norwegian Institute of Public Health, NO-0403 Oslo, Norway

2) Craniofacial Research, Murdoch Childrens Research Institute, Royal Children's Hospital, 3052 Parkville, Victoria, Australia

3) Department of Public Health and Primary Health Care, Faculty of Medicine and Dentistry, University of Bergen, NO-5018 Bergen, Norway

Correspondence: Håkon K. Gjessing, Norwegian Institute of Public Health, Division of Epidemiology, P.O. Box 4404 Nydalen, NO-0403 Oslo, Norway
E-mail: hakon.gjessing@fhi.no Telephone: +47 21 07 82 41

ABSTRACT

Health investigators routinely collect DNA and environmental data from study participants in order to assess the impact of genetic and environmental risk factors on an outcome of interest. When planning a study, alternate study designs are evaluated to minimize bias and achieve a large enough sample size from available resources. With the enormous volumes of high-quality biomedical data housed within its numerous biobanks, Norway is particularly well-suited to spearhead the investigation of a wide array of exposures and outcomes in a systematic manner. The rich array of longitudinal phenotypic data also permits an assessment of gene-environment-*timing* interactions. Maximizing the research potential inherent in Norwegian biobanks is the overarching aim of *Biobank Norway*, an infrastructure project recently funded by the Norwegian Research Council. The development of advanced statistical tools for the analysis of high-throughput genomic data is critical to fulfill this aim and biostatistics platforms have been key elements of major biobank harmonization initiatives. However, many of these approaches have focused on traditional case-control designs. To exploit the particular advantages inherent in the Norwegian Mother and Child Cohort Study (MoBa), we describe here models to analyze the special data configurations available with offspring-parent designs. These models and the statistical tools outlined in this review were developed through the support of *Biohealth Norway*, a biobank platform funded by the Norwegian Functional Genomics Research Program (FUGE).

INTRODUCTION AND BACKGROUND

Most epidemiologists have now become accustomed to modeling the effect of an allele or haplotype in just the same way as they would normally model an environmental exposure. Certain aspects of genetic variation are, however, intrinsically different from other types of exposure. Whereas genotypes are naturally randomized through meiosis and remain stable over time, environmental exposures may be subject to seasonal variations and lifestyle/behavioral changes. Nationwide prospective pregnancy and birth cohorts, such as the Norwegian Autism Birth Cohort (ABC) Study, are particularly attractive because they allow studies of GxE interaction and epigenetics in a temporal context (1,2). The current review focuses on a special study design that involves the collection of genetic and environmental data on a group of affected offspring (also termed "cases") and their biological parents, and a corresponding group of unaffected offspring and their biological parents (Figure 1). Collectively, we refer to these nuclear family collections as "offspring-parent triads" throughout this paper.

A prerequisite for the offspring-parent triad approach is that it must be possible to obtain DNA from the child's parents. This is not always possible if the disease in question is late-onset such that the offspring themselves are adults as, for example, is the case with Alzheimer's disease. However, if recruitment occurs through a child treatment center, for example a surgical unit treating babies with orofacial clefts, the parents will often be involved and present at the center, creating both opportunity and motivation to donate biological specimens. Even if DNA is not available from one or

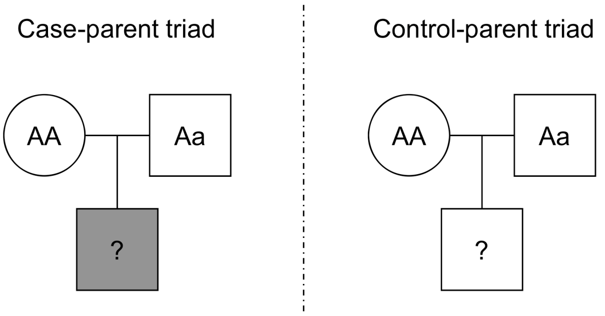

Figure 1. ***The "hybrid design" for family-based association analyses***. The hybrid design consists of a case-parent triad and a control-parent triad. The genotypes of the parents and offspring are known for a SNP at an autosomal locus, with *A* representing the common allele and *a* the variant allele. Under Mendelian transmission, the probability of an *AA* genotype is the same as that of *aA* in both the case-parent and control-parent triads. However, if the trait is associated with a particular genotype, its distribution among cases will differ from that expected under Mendelian transmission. The analysis consists of testing for this asymmetry.

more family members, data from incomplete triads can still be used by applying a statistical procedure that accounts for the missing genotypes. The offspring-parent triad design is therefore particularly well-suited for studies of birth defects such as orofacial clefts (3) and neural tube defects (4), and pregnancy-related conditions such as pre-eclampsia (5) and prematurity (6).

In this review, we use a large population-based study of orofacial clefts in Norway to illustrate the utility of offspring-parent triads in exploring different causal scenarios, including fetal and maternal gene-effects, gene-gene (GxG) interaction, and gene-environment (GxE) interaction. The paper includes a tutorial on how to perform the analyses using "Haplin" – a statistical software package specifically designed for analyzing genetic and environmental exposures in offspring-parent triads and case-control collections [Gjessing et al. (7)]. We apply a novel "hybrid design" that combines the merits of the case-control and offspring-parent triad designs. Not only does this hybrid design enhance statistical power by providing more controls per case, it also allows an estimation of the main effect of an exposure. Lastly, we discuss opportunities and challenges in using offspring-parent triads to analyze genome-wide data. Although the focus of this paper is on orofacial clefts, all the methods presented herein can easily be adapted to the study of other complex traits.

THE CASE-PARENT TRIAD DESIGN

Perhaps the most common approach to genetic association analysis is the standard case-control design, in which only case children and independent control children are used. Comparing allele frequencies of cases and controls can reveal loci associated with disease. An inherent danger of the case-control design, though, is *population stratification*, where marker allele frequencies vary across unrecognized subpopulations in the case and control groups, producing a spurious association between genotype and disease. However, the true impact of population substructure in well-designed case-control studies is up for debate (8,9).

One way to control for population stratification is to use only case-parent triads, i.e. genotyping case children and their parents, but leaving out the unaffected control-parent triads. Case-parent triads avoid the problem of population stratification by effectively using non-transmitted parental alleles as controls, to be compared with the alleles transmitted to the case child. In this setting, both "case" and "control" alleles derive from the same individuals and are thereby guaranteed to be selected from the same population subgroup. While not quite as (statistically) effective as the case-control design, the case-parent triad design allows an investigation of a range of causal scenarios with relatively high precision. These include fetal and maternal gene-effects, parent-of-origin effects, and the effects of GxG and GxE interaction. For instance, GxE interaction can be assessed simply by estimating the gene-effects in exposed and unexposed groups separately and then comparing the results.

Even for the case-parent triad design there are assumptions that should be met, however. It depends heavily on Mendelian transmission, so that to be valid, the probability of an offspring receiving any of the three possible genotypes from a particular parental mating type must follow Mendelian probabilities in the population (10). An example where this would be violated is when homozygosity for a variant allele increases the risk of early fetal death. If that is the case, live-born cases under study will appear to lack that particular variant, giving the impression that the genetic variant is related to the phenotype (11). In particular, this could lead to bias in studies of genes influencing preterm births, since very preterm births may be registered as spontaneous abortions, not preterm births.

For the investigation of maternal effects to be valid in the case-parent design, mating must be symmetric with regard to genotype. This means that the frequency of *Aa* mothers married to *AA* fathers should not differ significantly from the frequency of *Aa* fathers married to *AA* mothers. This assumption is needed because the distribution of the variant allele in mothers is compared to a null model in which the maternal and paternal allele counts are symmetric within each mating type (11).

For a valid estimation of GxE interaction, the genotype and environmental exposure must be independent, conditional on parental genotypes (10). In practice, this means that most situations where genes and environment are correlated in the population "at large" are unproblematic. Distortions may occur, however, if the genetic variant also influences an individual's tendency to be exposed, either through appetite or aversion (12,13). A good example is a person's aversion toward heavy alcohol-drinking, which appears to be correlated with a genetically-determined slower detoxification of alcohol (14).

THE HYBRID DESIGN

A notable limitation of the case-parent triad only design is its inability to assess the main effect of an environmental exposure. Comparing genetic effects in the exposed and unexposed triads will reveal interactions, but does not elucidate whether the environmental exposure is protective or harmful. In other words, the direction of the effect is not known. While the case-parent triad design protects against population stratification, the downside is lower efficiency than a case-control design. As a rule of thumb, a case-control children pair (two individuals to be genotyped) provides the same power as a full case-parent triad (three individuals to be genotyped).

As a consequence, various "hybrid designs" have been proposed to combine the merits of the case-parent triad and case-control design. The full hybrid design involves complete case-parent triads together with complete control-parent triads, not necessarily the same number of controls as cases. Truncated versions of the hybrid design have also been suggested, such as leaving out the control child (and only genotype his/her parents), leaving out the control father, or using case-mother dyads together with control-mother dyads (15-18).

In contrast to the case-parent triad design, the hybrid design allows the main effect of an exposure to be estimated because it involves independent controls. The controls in turn add more statistical power to the analyses. As a rule of thumb, a complete case-parent triad provides two transmitted case alleles and two non-transmitted control alleles. Adding a complete control-parent triad adds four independent control alleles, since the alleles carried by the control child are already present in his/her parents. Hence, a complete control-parent triad counts as two full controls (16) (Figure 1).

The hybrid design can also be used to estimate genetic effects for loci that exhibit deviations from Mendelian transmission. This is done by comparing the relative risk in the case-parent triads with the relative risk estimated from the control-parent triads. The control-parent triads should reveal the size of the deviation and thus serve as a baseline.

The implementation of the hybrid design in Haplin makes the standard "rare disease assumption", which allows relative risks and odds ratios to be used interchangeably. This is what enables the relative risk estimates from the case-parent triads to be combined with odds ratio estimates deriving from the case-control comparison. This assumption is reasonable for orofacial clefts, given the relatively low overall risks of CL/P and CPO.

While the hybrid design draws advantages from both the case-parent and the case-control designs, it is also to some extent influenced by population stratification. Since it incorporates a case-control component, the bias in the latter may creep into the overall estimate. Although the effect is lower than for the case-control design alone, it may still be noticeable.

SOFTWARE

A multitude of computer programs have been developed for statistical association analyses, in particular for the traditional case-control design. For case-parent triad analyses, the selection is somewhat more limited, and even more so regarding software that can handle hybrid designs. Most of the analyses described above can be done using Haplin – a statistical software specifically designed for analyzing genetic and environmental exposures in offspring-parent triads and case-control collections (7). It is based on log-linear modeling as originally described in (10-13,19-23), and was one of the first statistical software to extend the case-parent triad approach to consider haplotypes in a gene or region of interest. Although phase is not known from the observed SNP genotypes alone, Haplin can reconstruct haplotypes from the multi-SNP data and estimate the relative risk associated with a given haplotype. This is particularly advantageous over other competing methods that only provide a test of significance. Haplin is also equipped with an optimal imputation procedure to account for missing genotypes in a particular triad, providing additional flexibility if a SNP fails to be assayed in an individual or a parent chooses not to participate in the study.

Haplin is implemented in the publicly available ***R*** statistical package (24) and is freely downloadable from our web site at http://www.uib.no/smis/gjessing/genetics/software/haplin. A user-friendly graphical user interface (GUI), which includes some (but not all) of the Haplin functionalities, is also available at http://haplin.fhi.no.

A MAXIMUM LIKELIHOOD APPROACH TO ESTIMATION

Haplin implements a full maximum likelihood (ML) model for estimation. While other tests may be easier to implement and faster to estimate, the ML approach provides a full estimation framework. As a consequence, Haplin can compute explicit estimates of relative risks, with asymptotic standard errors and confidence intervals. A likelihood ratio test (LRT), Wald test, or score test can be used for contrasting two or more statistical models. In addition, the expectation maximization (EM) algorithm can be used for imputation, which consists of filling in genotype data that are missing at random (due to failed genotyping), reconstructing unknown haplotype phase, or even imputing data on family members missing "by design" – for instance if case-fathers were not available for genotyping.

At the heart of Haplin is a generalized linear model (*glm*) being estimated from the observed genotype frequencies. This is the M-step of the EM algorithm. The E-step consists of all three types of imputations, performed in a single step. The algorithm then alternates between the M-step and the E-step until convergence is achieved. The results obtained from the EM algorithm correspond to the maximum likelihood estimates of the model, which include gene frequencies and all relative risk parameters. However, to obtain correct standard errors, confidence intervals and LRT for the models, Haplin corrects for the fact that imputation has taken place. If the imputed data were used uncorrected, they would seem to contain more information than what is actually available in the raw data. This is adjusted for in the final output.

Finally, for studying GxE interaction, the Wald test is flexible in that it allows a comparison of any set of estimated parameters across two or more strata of exposure. The LRT can be useful as an additional check for model implementation and estimation.

MULTIPLE TESTING

In standard epidemiological analyses, multiple testing is rarely made an issue. Even if researchers in practice often "dig around" for anything apparently significant to report, this is usually not taken into account when publishing results. As a consequence, a large number of reported associations will be false positives that fail to replicate beyond the initial publication. In genetic epidemiology, this problem escalates to a degree where corrections must be performed, regardless of how tempting the initial results may be. As an illustration, a genome-wide association study (GWAS) will produce, say, 1 million different p-values, most of which will represent SNPs totally unrelated to disease. If a line for significance is drawn at, say, the customary 5%, then about 50,000 false positives will appear, completely obscuring any true positive results. It is clear that this situation is untenable, and appropriate measures must be taken.

The most obvious way to overcome the multiple-testing problem is to increase the sample size. With a large sample size, the true signals will generate p-values that are very small, for instance in the order of 10^{-7} or smaller. In that case, the line for significance can be set much lower, and the SNP that are still significant are likely to be true signals. Samples could also be used more effectively in, for instance, two-stage designs, where only the apparently significant SNPs from the first stage are genotyped and verified in the second stage.

The large number of SNPs available in a GWAS study means that the traditional statistical testing regime is less relevant, since *a priori* most associations would be expected to be negative. A more sophisticated approach is through the control of the false discovery rate (FDR) criterion proposed by Storey and co-workers (25). This leads to a set of "q-values" to replace the original p-values. The q-values thus relate to the false discovery rate, whereas p-values relate to the type I error in standard testing. In general, the multiple-testing problem can be approached through an empirical Bayes approach (26).

Another helpful tool is the quantile-quantile (QQ) plot, which can be used to inspect visually whether association analyses have produced more significant p-values than expected by chance. This represents a simpler alternative to using a full correction for multiple-testing. If none of the markers are associated with risk, the p-values in a QQ plot are expected to fall along the straight sloping line representing the null distribution. They would otherwise fall above this line at the most significant end of the QQ plot. In Haplin, the function `pQQ` is used to produce QQ plots.

OROFACIAL CLEFTS

As our prime example, we will study orofacial clefts, which is determined prenatally and possibly determined by fetal, maternal, and/or environmental effects. As a prelude to the tutorial, we first provide a brief introduction to orofacial clefts, followed by a description of the Norwegian Facial Clefts Study (NCL). More details on the study can be found at www.niehs.nih.gov/research/atniehs/labs/epi/studies/ncl/index.cfm.

Orofacial clefts: A common birth defect of complex etiology

Orofacial clefts include cleft lip (CL), cleft lip and palate (CLP) and cleft palate only (CPO). Because CL and CLP are regarded as variants of the same defect, only differing in severity, they are routinely lumped together to form the single group of cleft lip with or without cleft palate (CL/P). Collectively, these defects are the most common craniofacial birth defects in humans, affecting approximately 1/800 live births worldwide (3). Despite corrective surgery, patients experience a lifetime of functional, social and aesthetic challenges. The extensive medical and behavioral interventions needed to treat these defects impose a substantial economic and personal health burden which can persist from infancy to childhood and throughout life (28, 29). Not only are clefts the single most common craniofacial birth defects, they also appear to be associated with a higher risk of cancer in later life and an increased overall mortality well into adulthood (30-34).

The risk of recurrence of clefts is 30-40 times higher among those with an affected first-degree relative compared with the background population (35-38). A large Danish twin study recently reported heritability estimates of 91% for CL/P and 90% for CPO (36). The same study also found relatively small environmental contributions for either type of clefts, with 9% for CL/P and 10% for CPO respectively.

These and other related studies [reviewed in (3,39)] point to a very strong genetic component to clefting. Although the environmental contribution is likely to be smaller (36), assessing the joint impact of environmental risk factors and susceptibility alleles is important in solving the riddle of why some babies are born with clefts whereas the vast majority are not.

Study participants, candidate genes, and SNPs

We use offspring-parent triads collected in a nationwide case-control study of orofacial clefts in Norway (1996-2001). Mothers of babies with clefts were invited to participate in the study through the two surgical clinics appointed to treat all clefts in Norway. The participation rate was 88%, and 377 cases with CL/P and 196 cases with cleft palate (CPO) were recruited. During the same years, controls were randomly selected from all live births recorded in the Norwegian Medical Birth Registry. Of 1006 eligible control-mothers, 76% ($N = 763$) agreed to participate [see (40) for further detail]. Genotypes were available for 1536 SNPs in 357 candidate genes for orofacial cleft (41). The complete lists of genes and SNPs are provided in Supplementary Tables S1 and S2 of the original article (41).

For genome-wide analyses, we use a collection of case-parent triads from a recently conducted GWAS on orofacial clefts (42). These triads were recruited from seven Asian and six European/US recruitment sites (13 populations in total). Offspring-parent triads from the Norwegian study represent one of the European nodes in this international cleft consortium [details are provided in (42)].

TUTORIAL

In the next sections, we provide an informal tutorial on how to approach the various analytical issues discussed above, with practical examples using Haplin to analyze the clefts data. We describe each causal scenario in greater detail before providing practical examples. The main objectives are to estimate the effects of fetal and maternal genes, the effects of imprinting, X-linked genes, and the effects of GxG and GxE interactions.

ESTABLISHING GENE FREQUENCIES

Our model for risk estimation needs to establish the gene frequencies in the "background" population, i.e. the population at large, without reference to disease. This is used to provide a "baseline" for the gene frequencies, against which the gene frequency of, say, cases can be compared. In Haplin, SNP or haplotype frequencies are estimated as part of a full maximum likelihood model, jointly with genetic risk estimates, and reported as part of the output. Gene frequencies may be useful for instance in error-checking, for comparing a locus across populations, and for assessing the attributable risk associated with a particular locus. The following issues should be kept in mind, however:

- To reduce model complexity, Haplin typically assumes Hardy-Weinberg equilibrium (HWE), which in effect assumes that genotype frequencies for pairs of alleles can be computed as the product of the individual allele frequencies. This is a natural assumption in randomly mating populations, but may be more controversial in populations with substructures. Haplin makes an initial test for HWE at each locus before estimating the model.
- In data generated from standard SNP platforms, haplotypes are unphased and only SNPs are observed directly, not the full haplotypes. Haplin estimates haplotype frequencies by assuming HWE and imputing the unknown phase using the EM algorithm, in much the same way as missing genotypes are imputed.
- For a haplotype analysis to make sense, it is assumed that SNPs are in close linkage disequilibrium (LD), so that there is a low likelihood of recombination within a haplotype in one generation.
- With *K* SNPs, 2^K different haplotypes can in principle be constructed. With, say, 10 SNPs one could possibly observe 1024 different haplotypes, in various pairings within a genotype. In practice, only a handful of haplotypes will actually be observed in a population. Including too many SNPs in an analysis will only lead to too many possible haplotypes, with each haplotype being very unlikely. It hardly makes sense to include more than 5-7 SNPs in a haplotype analysis, depending on the degree of LD. A better approach is to use sliding windows of haplotypes, as described below.

All Haplin analyses incorporate an estimation of gene frequencies; no additional arguments need to be specified.

DESIGN, DATA STRUCTURE, AND MISSING DATA

Haplin analysis requires that the data are organized in a special format. For the case-parent triad data, Haplin uses three data columns for each marker, with genotypes for mother, father, and child, respectively, from left to right. A second marker should be placed to the right of the first marker, and so on so forth. Additional information, such as exposure variables, sex, and case-control status should be placed in separate columns to the left of the genetic data. The data structure is described in detail at our web site (http://www.uib.no/smis/gjessing/genetics/software/haplin/). To simplify data handling, the function **`pedToHaplin`** transforms data from a text file in the commonly used ped format into a text file in the Haplin format.

The default design in Haplin is the case-parent triad. It requires three columns for each marker and no additional columns need to be supplied. Missing genotypes are specified with an "NA" in place of the genotype. For case-mother dyads, i.e. when the father's genotype is not available, all genotypes for the father should be set to "NA". The standard design is specified with **`design = "triad"`**, but for the default design this is not necessary.

If the design includes independent control-parent triads, the data are placed in the same columns as the case-parent triad data. In addition, a column with 0 (controls) and 1 (cases) is placed to the left of the genetic data to identify cases and controls. The design is specified with **`design = "cc.triad"`**. This represents the complete "hybrid" design. If only control-children are available, not control-parents, columns for the parents should be included, but set to "NA" for missing data. The design is still specified as **`"cc.triad"`**. The same approach should be followed for two other recommended hybrid designs (16,18), where only control-mother dyads or only control parents are available.

Finally, if the design is the standard case-control design, with no parental information, the data file should contain one column for the genotype at each marker. To the left of the genetic data, there should be at least a column with the case-control information. In this case, the design is specified simply as **`design = "cc"`**. Note that if genotypes of both case-mothers and control-mothers are available, the analysis is commonly done by using only the maternal data with a **`"cc"`** design. However, if fetal (and perhaps also paternal) genotype data are available, it is recommended to use the hybrid design as described above. This allows correcting any maternal genetic effects for possible fetal effects of the same genes.

An X-chromosome analysis requires gender of the child to be specified, using a separate column with 1 (males) and 2 (females). In addition, the argument `xchrom = T` must be set. If the data file uses a case-control variable and/or a sex variable, their column positions (counted from left to right) in the data file must be specified using, for instance, `ccvar = 1, sex = 2`, to specify the first and second columns, respectively. In addition, `n.vars = 3` is used to tell Haplin that there is (for instance) three columns of data to the left of the genetic data.

Haplin uses an EM-algorithm to impute data, whether they are randomly missing genotypes, missing phase information for the haplotypes, or data missing "by design" (for instance if fathers have not been genotyped). To instruct Haplin to impute data, set the argument `use.missing = T`. Note that imputation is by default turned off. Families or SNPs with a large proportion of missing data should preferably be removed before analysis, since Haplin spends a fairly large amount of computational power to impute such data.

FETAL GENE-EFFECTS

The most obvious question in a genetic study of clefts is whether the genes of the fetus directly influence the risk. The risk associated with a single locus can be estimated as an increased risk of clefts for one genotype relative to the others. The actual risk of being born with clefts cannot be computed from the offspring-parent triad design alone without knowing the background population from which the cases are selected. However, the relative risks obtained when comparing one genotype to another can be computed both from the pure case-parent triad design and clearly also from the hybrid design which involves additional independent controls.

As a concrete example, when looking at a single SNP, say with major allele C and minor allele T, there are three possible genotypes: CC, CT, and TT. If CC is chosen as the reference genotype, one can estimate relative risks RR_{CT} and RR_{TT} associated with genotypes CT and TT, respectively. If the T variant increases the risk, these relative risks will typically be larger than 1.0, whereas if C increases risk they will typically be less than 1.0. Haplin allows estimation of both RRs separately, or with the assumption of a dose-response model where $RR_{TT} = RR_{CT}^2$, i.e. a multiplicative risk model.

Looking at a locus with multiple alleles, such as haplotypes over a set of SNPs, the model is more complicated since haplotypes combine in pairs, and with, say, 7 different haplotypes at a locus, it is possible to form 7(7+1)/2=28 different pairs of haplotypes as a genotype. Some of these pairs will be virtually non-existent. To reduce the modeling complexity, Haplin estimates two relative risks associated with each haplotype, one for a single dose and one for a double dose of that haplotype. As default, a haplotype is compared to the "average" of the other haplotypes, so that all other haplotypes serve as a combined reference category. This is referred to as a "reciprocal reference". As for the single SNP model, the RR parameters for multiple haplotypes can be assumed multiplicative, i.e. to follow a dose-response model.

A dose-response model in Haplin is specified through the argument `response = “mult”`. By default Haplin will estimate single and double doses separately, but assuming a multiplicative dose-response model may increase power by reducing the number of parameters to be estimated.

Haplin example

For a pure case-parent triad design, the simplest model in Haplin can be estimated with the command:

```
haplin(“C:/work/data.dat”)
```

Only the file name needs specification, all other arguments are chosen as default. If the data file contains more than one marker, the relevant marker (SNP) can be specified using:

```
haplin(“C:/work/data.dat”, marker = 2)
```

If no marker is specified in a multi-marker file, Haplin automatically builds haplotypes from all the available markers. Note that this may become too computer-intensive (and also meaningless) if too many markers are included, for the reasons discussed above. If data is missing from the file, typically when genotyping has failed or a family member has not been genotyped, missing data can be imputed using the `use.missing` argument:

```
haplin(“C:/work/data.dat”, marker = 2,
use.missing = T)
```

When Haplin runs, the first output is summary information on data and markers, followed by more detailed estimation results. Below is an excerpt of the analysis output for marker 1 for the interferon regulatory 6 (*IRF6*) gene, where we have used the Norwegian data on orofacial clefts:

```
haplin(“C:/work/data.dat”, marker = 1,
response = “mult”, use.missing = T, design
= “cc.triad”, n.vars = 7, ccvar = 2)
```

```
----Data summary:----
There were 17 rows with missing data
All rows retained in analysis

Number of triads in original file: 114

Accounting for possible loss of triads:
 Cause of loss  Triads removed  Triads remaining
 Missing data                0               114
 Mendelian incons.           0               114

Triads remaining for analysis: 114
```

Haplin has thus found 114 triads in the file, of which 17 contained missing data. However, all triads were retained in the analysis, with missing data imputed. There were no Mendelian inconsistencies. The output then continues:

```
----Estimation results:----
Number of haplotypes: 2
```

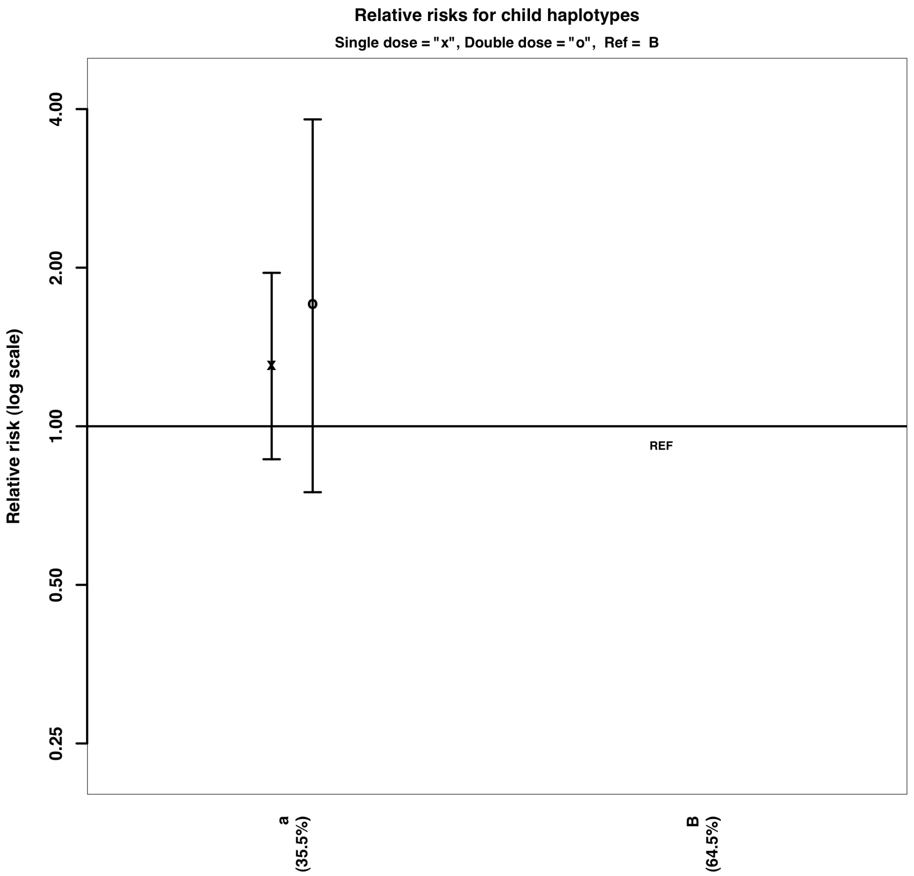

Figure 2. ***Fetal gene-effects***. Estimated relative risks of isolated cleft palate only (CPO) among 114 babies carrying a single or double dose of the variant allele *a* versus carrying none (allele *B* is the reference). Vertical bars represent 95% confidence intervals, shown on a logarithmic scale.

```
Haplotype frequencies with 95%
confidence intervals:
 Haplotype  Frequency(%)  lower  upper
 a           35.5         29.3   42.4
 B           64.5         57.6   70.7
```

There are two "haplotypes" in a single SNP, here generically coded as "a" and "B" (the uppercase "B" is used to designate the most frequent allele). Then the relative risk estimates follow:

```
Single and double dose effects (Relative Risk,
RR) with 95% confidence intervals:

----Child haplotypes----
 Haplotype  Dose  RR    Lower CI  Upper CI  P-value
 a          S     1.31  0.866      1.96      0.201
 a          D     1.71  0.75       3.82      0.201

 B          S     REF
 B          D     REF
```

Allele 'B' is here chosen as reference. Having a single dose (S) of allele "a" thus increases the risk of clefts 1.31-fold. Note that this particular analysis assumes a dose-response relationship, which means that the relative risk associated with a double dose (D) of "a" equals $1.31^2 = 1.71$. This property is also reflected in a common p-value, 0.201, for both single and double dose. The result is non-significant (Figure 2).

MATERNAL GENE-EFFECTS

It is important to consider the role of maternal genetic factors when assessing risk, because the mother's genotypes partially controls the *in utero* environment of the developing fetus (7,11,12,23). Studies in animal models have demonstrated an ability of maternal gene products to directly intervene and protect the fetus. Specifically, Letterio et al. (43) showed that maternal Tgfb1 was able to cross the placenta and rescue *Tgfb1*$^{-/-}$ mice. Similar observations were made in an earlier experiment that tested whether maternal epidermal growth factor (Egf) could be transported to the fetus via the placenta (44).

The offspring-parent triad design allows a straightforward modeling of both maternal and fetal gene-effects without confounding from one another (41,45-52). There are several ways in which a variant allele may increase risk [adapted from (53)]:

- The variant allele increases risk *only* if carried by the fetus (contributing to a "fetal gene-effect"). As the variant allele in the parents does not contribute to risk, it will be over-represented in cases compared to the biological parents.
- The variant allele increases risk *only* if carried by the mother (contributing to a "maternal gene-effect").

The variant allele will be over-represented in case mothers compared with the case fathers.

- The variant allele increases risk *both* when carried by the fetus and by the mother. The model in Haplin then assumes that the relative risks for the fetal and maternal contributions can be multiplied together to obtain the joint risk of disease. This is akin to a logistic regression where several exposures can be included and adjusted for one another, and allows estimating both fetal and maternal risk when controlling for possible confounding with one another.

Haplin example

Including maternal effects in the Haplin analysis is simple. The argument `maternal` should be set to true, otherwise the analysis is unchanged:

```
haplin("C:/work/data.dat", maternal = T)
```

Below are the relative risk estimates for the same analysis as in the previous example (marker 1 of the *IRF6* gene), with the difference that maternal effects are included, and there is no assumption about a dose-response relationship:

```
haplin("C:/work/data.dat", marker = 1,
use.missing = T, design = "cc.triad",
n.vars = 7, ccvar = 2, maternal = T)
```

```
----Child haplotypes----
 Haplotype  Dose  RR    Lower CI   Upper CI   P-value
 a          S     1.88  1.13       3.15       0.0182
 a          D     1.32  0.546      3.12       0.542

 B          S     REF
 B          D     REF

----Maternal haplotypes----
 Haplotype  Dose  RR    Lower CI   Upper CI   P-value
 a          S     1.27  0.802      2.04       0.311
 a          D     0.848 0.386      1.84       0.688

 B          S     REF
 B          D     REF
```

Again, it appears that a single dose of "a" in the fetus increases the risk of clefting, this time as much as 1.88, and with a borderline significance. A double dose does not seem to increase the risk any further, so the effect of "a" appears to be more dominant than dose-response. However, it should be kept in mind that the separate double dose relative risk estimates are often unstable if the minor allele frequency is exceedingly low or the sample size is too small. This is reflected in the wide confidence interval for the double dose estimate. In this example, maternal alleles do not seem to have any significant effect.

In addition to the text output, Haplin also produces a plot of the relative risks for both fetal and maternal effects (Figure 3).

EFFECTS OF X-LINKED GENES

Most association analyses have primarily targeted autosomal markers, most likely because the vast majority of statistical methods for association analysis were originally designed for autosomal markers. The discovery that genetic variants on the X-chromosome may be associated with several complex traits has prompted the development of a variety of statistical methods for analysis of X-linked markers. The majority of these methods are extensions of the transmission/disequilibrium test (TDT) and include the following: (i) X-linked sibling TDT (XS-TDT) (54); (ii) reconstruction-combined TDT for X-chromosome markers (XRC-TDT) (55); (iii) X-linkage TDT test (X-TDT) (56); and (iv) X-chromosome pedigree disequilibrium test (XPDT) (57). Two other tests, the "association in the presence of linkage (APL) test that accommodates X-chromosome markers" (X-APL) (58) and the "X-linked quantitative trait loci linkage mapping" (X-QTL) (59), are based on comparing observed *vs.* expected distributions of a specific allele or haplotype in affected siblings, conditional on the parental genotypes. Because these methods are based on the TDT, they can only provide a p-value for association and not estimates of genetic risk. One exception is the likelihood ratio test (LRT) of association for X-linked markers (X-LRT) (60).

We have implemented a new functionality in Haplin to enable association analysis of X-linked markers. The model is similar to the X-LRT approach, but we extend the model to haplotypes and a selection of gene-effect models. We can test five separate models as outlined in Table 1, assuming common or different parameters for boys and girls:

- **Model 1**: Common baseline risk, common relative risk, no X-inactivation (1 parameter to be estimated)
- **Model 2**: Different baseline risks, common relative risk, no X-inactivation (2 parameters)
- **Model 3**: Different baseline risks, different relative risks, no X-inactivation (3 parameters)
- **Model 4**: Different baseline risks, common relative risk, X-inactivation (2 parameters)
- **Model 5**: Different baseline risks, different relative risks, X-inactivation. This is the "free model" with 4 parameters to be estimated.

Note that more parameters require fewer assumptions, but fewer parameters result in higher power (assuming the model is correct). In Models 4 and 5, we can account for X-chromosome inactivation, a mechanism by which one of the two copies of the X-chromosome in females is inactivated to ensure similar gene-dosage between the two sexes. Hence, we model the relative risk for males with one X_2-allele as being equal to that of an X_2X_2 homozygous female (Table 1). We present here an example that implements Model 4. Different baseline risks are appropriate in this case given the higher prevalence of CL/P among males and the higher prevalence of CPO among females.

Haplin example

To do an analysis of markers on the X-chromosome in Haplin, one can use a command like the following:

```
haplin("C:/work/data.dat", xchrom = T,
nvars = 3, sex = 2, use.missing = T)
```

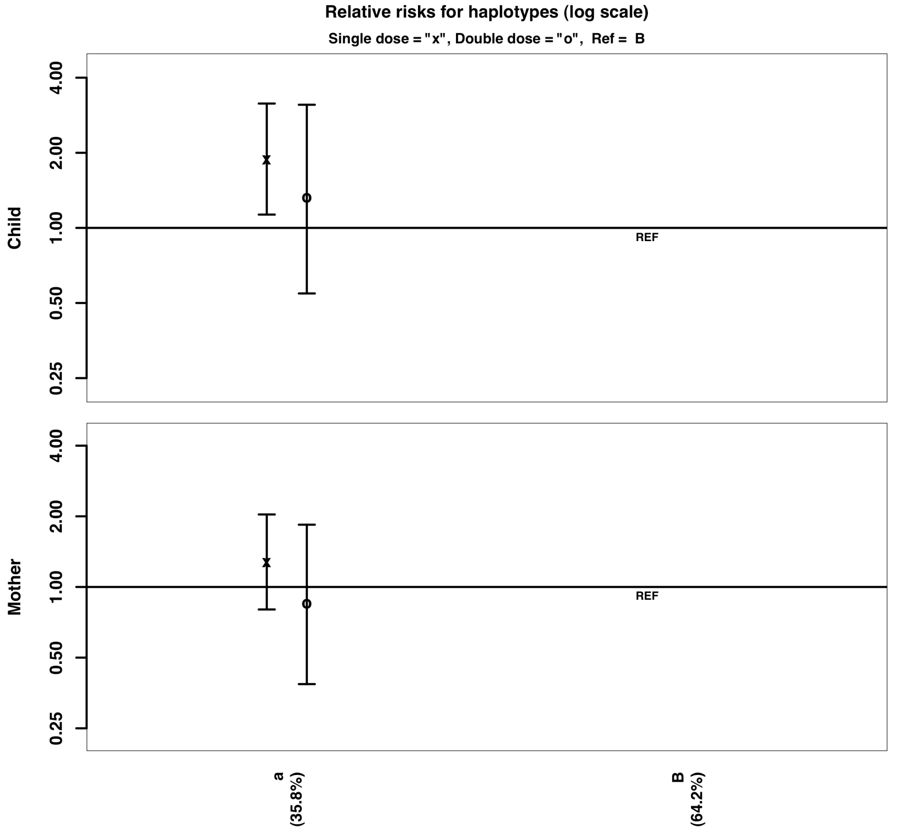

Figure 3. ***Fetal and maternal gene-effects.*** The upper half of the panel shows relative risk estimates based on the baby carrying the variant allele *a*; the lower half shows the corresponding estimates when the mother carries the variant allele *a*. Vertical bars represent 95% confidence intervals, shown on a logarithmic scale.

Table 1. Assorted parameterization models for X-linked gene analysis.

	Male case		Female case		
Model	X_1	X_2	X_1X_1	X_1X_2	X_2X_2
Model 1	B	B * RR	B	B * RR	B * RR^2
Model 2	B_M	B_M * RR	B_F	B_F * RR	B_F * RR^2
Model 3	B_M	B_M * RR_M	B_F	B_F * RR_F	B_F * $RR_F{}^2$
Model 4	B_M	B_M * RR	B_F	1/2 * B_F * (1 + RR)	B_F * RR
Model 5	B_M	B_M * RR_M	B_F	B_F * RR_{F1}	B_F * RR_{F2}

$\mathbf{X_1}$ denotes the common allele and $\mathbf{X_2}$ the variant allele at the SNP in a given gene. '*' denotes the product term; **B** represents the common baseline risk for males and females; $\mathbf{B_M}$ is baseline risk for males only; $\mathbf{B_F}$ is baseline risk for females only; **RR** is the common relative risk for males and females; $\mathbf{RR_M}$ is the relative risk for males only; and $\mathbf{RR_F}$ is the relative risk for females only. In Model 4, the risk for an X_1X_2 female will be an average of the two homozygotes, i.e. $(B_F + B_F{*}RR)/2 = B_F(1 + RR)/2$. Strictly speaking, this is not a log-linear model, so Haplin replaces the heterozygous risk with $B_F\sqrt{RR}$, i.e. the geometric mean of the two homozygous risks. Note also that only the ratio of the baseline risks $\mathbf{B_F}/\mathbf{B_M}$ can be estimated in the case-parent triad setting.

Parent-of-origin effects and effects of imprinted genes

We can have a parent-of-origin effect if transmission distortion to affected offspring is stronger for mothers than for fathers (19). As noted above, the mother can influence the development of the fetus through the action of her own genes and through providing the prenatal environment for the fetus. The mother's genetic contributions may be nuclear (genomic imprinting) or extra-nuclear (mitochondrial inheritance) (61). Genomic imprinting refers to the situation where certain genes are differentially expressed according to whether they are inherited from the father or the mother (62). This means that although genes from both the mother and father are present in the embryo, they do not operate at the same level.

The effect of imprinting ranges from the total inactivation of a gene to its reduced expression in specific tissues. Even though the mechanisms underlying genomic imprinting in mammals are still not completely understood, they are thought to involve DNA methylation of cytosine-rich segments. Studies in mice have identified an array of genes whose expression are restricted to either the maternal or paternal allele (62), and there is some evidence to suggest that genomic imprinting may also control intrauterine embryonic growth in humans. Notably, paternally-expressed genes appear to be involved in placental development, whereas maternally-expressed genes appear to influence embryonic growth (62). It is therefore important to separate maternal gene-effects from the effects of imprinting (22).

The offspring-parent triad study design can easily incorporate tests for parent-of-origin effects (19). To test for the effects of imprinting in offspring-parent triads, one simply compares the proportion of children who receive a copy of the variant allele from the mother versus the father (19). A statistically significant difference in these proportions does not mean biological proof of genomic imprinting. Appropriate biological assays are needed to substantiate the statistical findings.

Haplin example

In Haplin, parent-of-origin effects can be estimated by comparing the relative risk associated with the allele transmitted from the mother with that transmitted from the father. The currently distributed version of Haplin does not include this functionality, but it is likely to be included in an upcoming version.

GENE-GENE (GXG) INTERACTION

Segregation analyses have shown that orofacial clefts are unlikely to be caused by a single gene (63), but rather through the complex interplay of multiple genes and environmental factors, each making a minor contribution to the overall risk. Traditional linkage-mapping techniques would therefore face serious challenges under this setting. Similarly, if a few major genes were to contribute to the risk of clefts, the presence of other modifying genes that exert small effects on the major genes would render linkage mapping intractable.

Association studies on the other hand are better powered than linkage approaches to detect variants of more modest effects, provided that the genetic marker is close enough to exhibit strong linkage disequilibrium (LD) (64,65). Box I details the main differences between linkage and association approaches, while Box II describes the concept of linkage disequilibrium (LD) and its importance in allelic association studies. It should be noted, however, that GxG analyses can only test for a risk interaction, as opposed to a specific interaction between gene products in a given biological pathway (66).

Finally, it is worth noting that pathway analyses are useful in narrowing down the search for significant signals by focusing only on SNPs or genes within a relevant biological pathway, an approach that also allows for more targeted GxG interaction analyses. Examples from our own work include investigations of the renin (*REN*) and angiotensinogen (*AGT*) genes in the RAS system that controls blood pressure (67), and of genes that metabolize folic acid in the folate pathway (68). Pathway-specific databases such as KEGG (http://www.genome.jp/kegg/pathway.html), PANTHER (http://www.pantherdb.org/pathway), INGENUITY Pathways Analysis (IPA; http://www.ingenuity.com) or BIOINCARTA (http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways) are useful in identifying genes sharing a common biological pathway.

Haplin example

The reader is referred to Vefring et al. (5) in which we assessed the joint effect of maternal *AGT* and fetal *REN* haplotypes. Briefly, Haplin was first used to identify maternal *AGT* haplotypes that could be associated with the risk of preeclampsia. Next, the preeclampsia triads were stratified by presence/absence of the maternal *AGT* risk-haplotype to determine whether the effect of fetal *REN* haplotypes differed significantly across strata of maternal *AGT* haplotype. A LRT was used to test for such a difference.

GENE-ENVIRONMENT (GXE) INTERACTION

It has long been hypothesized that orofacial clefts result from the complex interplay of multiple genes and environmental factors, but only recently have practical approaches become available for robust investigation of this hypothesis (3,39). The main rationale for investigating GxE interaction is to determine the potential for public health intervention on environmental factor(s) which alone could reduce the occurrence (and recurrence) of the disorder, particularly in genetically susceptible subgroups of the population. This rationale is supported by findings in animal models. In mice, for example, the spontaneous clefting rate among the cleft-susceptible CL/Fr strain is about 20% compared to less than 10% in the normal C57BL/6J strain. However, this rate can easily be increased to almost 100% at certain dosages of 6-aminonicotinamide (a vitamin B3 inhibitor) (69). Just as some mouse strains are more susceptible to external teratogens (69,70), human fetuses carrying specific high-risk alleles may be more sensitive to particular teratogenic agents.

Although the role of environmental factors in orofacial clefting is well recognized, only a few GxE interactions have thus far been identified in human studies (71). Differences in exposure assessment, limited sample size, and study heterogeneity are among the known challenges in studies of GxE interaction (66,72,73). A cleft study therefore needs to be sufficiently large and phenotypically well-characterized to provide the level of statistical power necessary to tease out the effects of GxE interaction (74). Although the difference in statistical power between the case-parent triad and case-control designs is generally small for gene-association studies (75), the case-parent triad design has superior power for GxE interaction studies (72).

Box I. Linkage *vs.* association analysis

Linkage analysis

Linkage analysis maps disease genes by looking for co-segregation of a marker allele with the disease among related subjects. If the marker allele and the disease-causing allele are located near each other on the same chromosome, they are highly likely to be transmitted together, because multiple crossovers between closely linked loci are very rare. The LOD score measures the strength of evidence in favor of linkage. It is the log to base 10 of the ratio of the likelihood of the marker and the disease gene being linked (recombination fraction $\theta < 0.5$) compared with the likelihood of no linkage (i.e., $\theta = 0.5$). The threshold for accepting linkage is a LOD score of +3 (a likelihood ratio of 1:1000) and that for exclusion is –2.

Association analysis

Association is a statistical term that describes the co-occurrence of two investigated factors significantly more often than what would be expected based on chance alone. A marker allele is said to be associated with a disease if its frequency is significantly higher (or lower) among affected individuals compared to predicted values from the general population. Two types of association studies are frequently used: *population-based* and *family-based*. The population-based approach relies on a standard case-control design where marker allele frequencies are compared between a set of unrelated affected individuals and a set of matched controls. By contrast, the family-based approach tests for the asymmetric distribution of a particular target allele among affected offspring and their biological parents. A positive association has several possible interpretations: (i) the marker allele itself is the disease-causing allele, (ii) the marker allele is in linkage disequilibrium (LD) with the disease-causing allele, (iii) the association is spurious due to population stratification (the population contains several genetically distinct subsets), or (iv) the association is merely a Type I error (a false positive).

How does linkage differ from association?

An association can have many causes, not all of which are genetic. Linkage by contrast is a specific genetic relationship between loci (not alleles or phenotypes). If linkage exists between a marker locus and a disease, co-segregation will be observed within a family irrespective of which allele at the marker locus is under scrutiny. Linkage looks at within-family differences between marker alleles and the disease, whereas allelic association studies exploit across-family associations. However, in populations that share a substantial number of common ancestors, linkage and association will tend to converge.

Box II. Linkage disequilibrium (LD)

LD in association mapping

LD occurs when a particular marker allele is so close to the disease allele that they are less likely to be separated by recombination. They are thus co-inherited over many generations. This represents a deviation from Mendel’s second law of independent assortment of genes. Under Hardy-Weinberg equilibrium, the frequency of a two-locus haplotype is the simple product of their individual frequencies. If, however, the population frequency of the haplotype is either in excess or in deficiency, the two loci are said to be in LD. LD will create an association if a significant proportion of the disease chromosomes derive from one not too distant common ancestor. Since the number of meiotic events observed in linkage studies is considerably less than that in LD studies, very dense maps of markers are usually needed for LD-based gene mapping. LD rarely extends more than one centimorgan (cM) from a susceptibility locus, unless the study is being conducted in population isolates in which genetic variability is often reduced. Even though allelic association may potentially outperform linkage in detecting weak susceptibility alleles, the decrease in power with genetic distance is much more dramatic in allelic association studies. When a specific genetic variant under study is not in the actual disease-causing gene, allelic association will nevertheless occur due to LD. This is more the rule than the exception in studies of complex diseases, where allelic heterogeneity will typically tend to limit the strength of an association.

Factors affecting LD

A large number of factors that are highly stochastic in nature interact to create substantial variation in the strength of LD across different populations and even across different segments of the human genome. These include founder population size, mutation rate, gene conversion, recombination, and natural selection. LD decays largely according to the recombination distance between markers and the number of generations that have elapsed since the susceptibility allele was first introduced into the population. This is mathematically represented by $D_t = D_0(1-\theta)^t$, where D_0 is the amount of LD at time 0, θ is the recombination fraction, and t is the number of generations that have elapsed. LD around rare alleles is thus expected to have longer range since such alleles are generally young and have undergone lesser reshuffling by recombination. Demographic factors such as inbreeding and population structure inflate LD. Inbreeding increases LD in a population by reducing the level of diversity. The same applies to recently admixed populations where a difference in allele frequency contributes to LD. Other factors that inflate LD include population bottlenecks (resulting in fewer founders, and hence lesser diversity) and natural selection where certain alleles that confer a selective advantage are swept to “fixation”.

A risk-conferring allele can interact with an environmental exposure in the following manner [adapted from (53)]:

- The variant allele increases risk *only* when carried by the fetus and, at the same time, the fetus is exposed to the environmental agent (e.g. maternal smoking or alcohol intake during the 1st-trimester of pregnancy). Here, we expect to observe a positive interactive effect between the child’s genotype and the environmental exposure.
- The variant allele increases risk *only* when carried by the mother and, at the same times, she is exposed to the environmental agent. In this case, we expect an interactive effect between the mother’s environmental exposure and her genotype.
- Mixed scenarios of the above.

Case-parent triads have previously been used to investigate GxE interaction in clefting (76-78), and different extensions of this method have been proposed (79-81). In a case-parent triad setting, GxE interaction is assessed by comparing the transmission of a risk-allele or risk-haplotype to affected offspring in triads of exposed *vs.* unexposed mothers. A statistically significant difference between the two transmissions would suggest a multiplicative interaction. To test whether the relative risk estimates are significantly different between the two strata of exposed and unexposed mothers at each locus, two approaches can be used, each with its own set of advantages/disadvantages. The first approach employs dummy variables for exposure categories in the design matrix of the log-linear model to create the appropriate interaction terms. Interaction terms for haplotype frequencies can also be added, allowing for possibly different haplotype frequencies over the different exposure categories. This approach is advantageous because it maximizes statistical power and seamlessly integrates the GxE interaction analysis with the already existing maximum likelihood model framework in Haplin. The drawback is that the design matrix of the log-linear model may become prohibitively large, in particular when dealing with multiple haplotypes and a large amount of missing data.

The second approach uses "post-estimation". In this case, Haplin can be run on separate data files, one for each exposure category. The results from each exposure category can then be tested against one another using either a Wald test or a score test. The Wald test has the advantage of being computationally simple; only the parameter estimates and their variance-covariance matrix need to be stored. The score test is computationally more intensive because individual score contributions must be stored. However, the score test is readily extended to a correction for multiple-testing within a region, for instance, for all SNPs within a gene. Generally, the post-estimation strategy is computationally less intensive and allows the exact hypothesis to be specified and tested after the time-consuming part of the estimation has been performed, avoiding unnecessary re-runs of the full analysis.

Haplin example

To use post-estimation to test GxE interactions with Haplin, two steps must be followed. First, Haplin is used to estimate gene-effects in each stratum of the exposure covariate, using a command like this:

```
res.strat <- haplinStrat("C:/work/data.dat",
n.vars = 7, response =  "mult", markers = 1,
covar = 7, use.missing = TRUE,
design = "cc.triad", ccvar = 2)
```

Note that the exposure covariate (in the seventh column) is specified using `covar = 7`, and the result is saved in an ***R*** list called `res.strat`. Second, results from all strata are compared using the posttest:

```
posttest(res.strat)
```

The result will be the chi-squared test statistic and the corresponding p-value of the GxE interaction between the first SNP (`markers = 1`) and a given environmental covariate (e.g. smoking status).

USING OFFSPRING-PARENT TRIADS TO ANALYZE GENOME-WIDE DATA

Major advances in high-density SNP genotyping arrays have heralded a new era of gene discovery for complex traits. Approximately 0.5-2.1 million SNPs can now be interrogated in a GWAS, providing an unprecedented level of marker resolution for association mapping. Being an agnostic method, GWAS can potentially identify new disease-related genes and genetic pathways, providing a deeper insight into the pathogenesis of the disease compared to candidate-gene based efforts (82,83). Despite the popularity of GWAS, however, its utility in the clinic has been debated (84-86). This is because the vast majority of the GWAS findings explain only minor fractions of the overall phenotypic variance attributable to additive genetic factors (85,87). Even when positive GWAS signals are identified, it is increasingly recognized that the complex structure of association signals makes it difficult to identify the one or more etiologic variants contributing to those signals. This arises both because the signals may be acting over many hundreds of Kb of DNA, creating a huge pool of candidate SNPs to evaluate, and also because the variants do not arise in more recognizable DNA sequences such as coding or promoter sequences, in which we understand causality. In addition, the contributions of rare variants, identifiable only by linkage and/or deep-sequencing approaches, are also likely to be greater than expected, but there remain major challenges of strategy and cost to characterizing these variants. Other critical variables such as epigenetics and the environment have also not been easily incorporated into analysis.

In addition to the points noted above, the large number of tests and small odds ratios associated with risk alleles require very large sample sizes, often requiring collaboration between different research groups for a subsequent meta-analysis of the GWAS findings. This entails a strict adherence to harmonization protocols for the establishment of a reliable platform for sharing both genotype and phenotype information across participating cohorts (88). The first GWAS on orofacial clefts, performed in individuals of Central European ancestry, identified a susceptibility locus on chromosome 8q24 (89). This locus was subsequently replicated in three independent GWAS (42,90,91).

Haplin includes the function `haplinSlide`, which automates the analysis of a long sequence of single SNPs, or alternatively a sequence of overlapping sliding windows with haplotypes of length, for instance, equal to 4. Overlapping sliding windows will in principle increase the chance of "bracketing" a point mutation by having a haplotype with SNPs on each side of the muta-

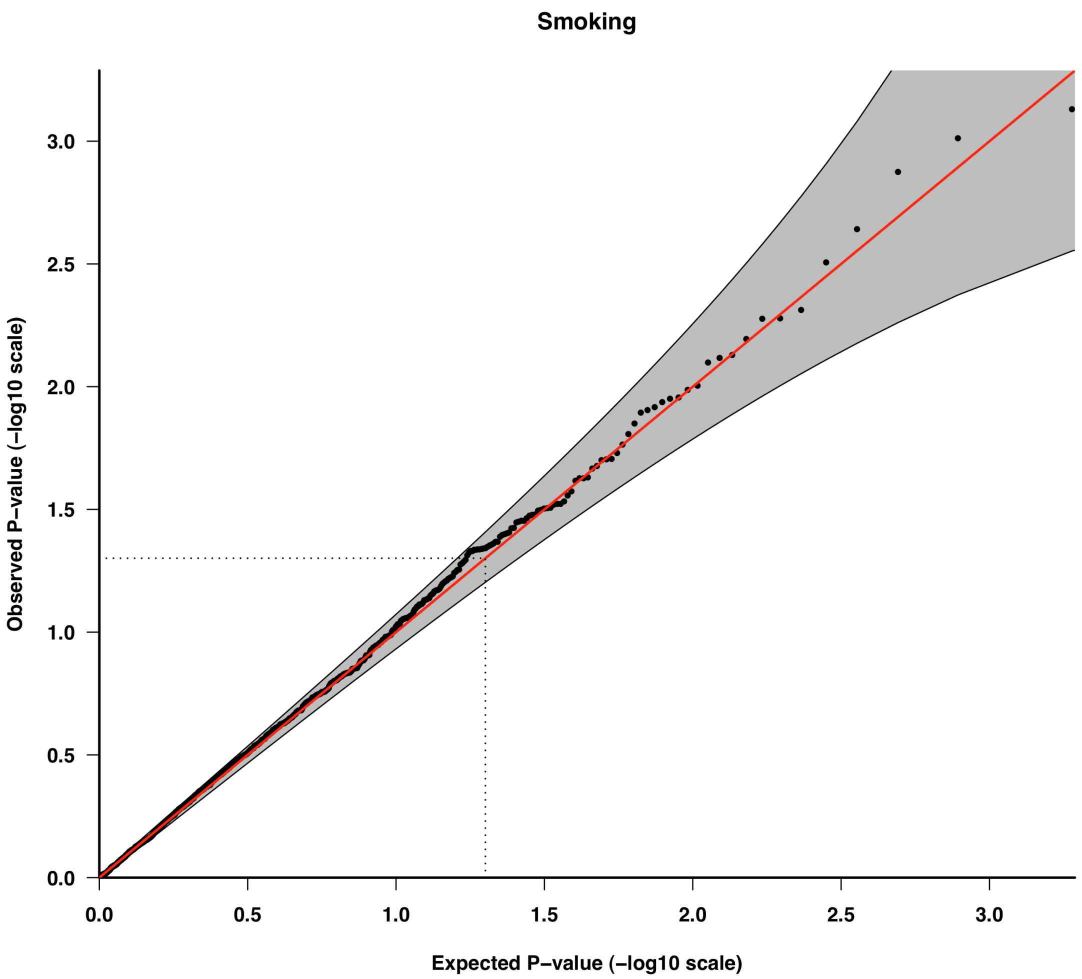

Figure 4. ***GxE interaction analysis***. The quantile-quantile (QQ) plot shows the p-values for interaction between maternal first-trimester smoking and isolated cleft palate only (CPO) for 1315 SNPs in 334 autosomal cleft candidate genes. The QQ-plot compares p-values (-log10 scale) with an expected uniform distribution under the null (sloping line). The pointwise 95% confidence bounds for the p-values are indicated by the grey shadings around the expected p-values. The stripped line in the QQ-plot indicates a p-value of 0.05.

tion. However, estimating haplotypes entails a certain loss of power due to the higher number of alleles taken into account, and the unknown phase of the haplotypes. It is *a priori* not obvious whether a single SNP approach or a sliding-window haplotype approach will have the best chance of detecting an association, so doing both might be worthwhile.

When analyzing SNPs in strong LD, and in particular when analyzing overlapping windows of length 4 haplotypes (which will contain three of the same SNPs as the neighboring haplotype), there is a strong correlation between results obtained from nearby SNPs or windows. Haplin has the possibility of computing a single summary p-value for a genetic region (such as a gene), based on all SNPs/windows within that region, corrected for the dependencies.

Haplin example

As described above, the main approach in Haplin to do a scan over a range of SNPs, either one at a time or in sliding windows, is through the function **`haplinSlide`**. The syntax for **`haplinSlide`** is almost identical to that of **`haplin`** itself, with the additional specification of window length (default is 1, meaning one SNP at a time). Output from **`haplinSlide`** is a list, each element of which is the result from running **`haplin`** on one window (or single SNP). Since the output from **`haplin`** can be large (there is a lot of information stored in the background), **`haplinSlide`** has an option to produce a tabular output, which corresponds roughly to the screen output from **`haplin`**, organized in a table.

Moderately large data files with, say, a few thousand SNPs, can be read directly into **`haplinSlide`** as an ordinary Haplin data file. However, this is slow in the currently distributed Haplin (version 3.5), and there is only limited functionality for handling very large data sets such as those from GWAS studies without breaking them up "manually". However, the current beta version of Haplin includes this functionality, and it is due to appear in the next release. The new version is faster on moderately large files. In addition, it utilizes the data structure of the GenABEL library for handling GWAS files (92).

Having completed a scan by Haplin, p-values can be assessed using the function pQQ to produce a QQ-plot. As an example, Figure 4 shows p-values for the interaction between maternal first-trimester smoking and isolated CPO for the 1315 SNPs in 334 autosomal cleft candidate genes in the Norwegian candidate-gene study. None of the SNPs show significant departures from the null hypothesis of no GxE interaction.

CONCLUDING REMARKS

Dissecting the causal architecture of common, complex diseases will depend critically on large-scale pooling of biomedical data from different biobanks. With its strong tradition of centralized health registries, Norway is in a prime position to explore the scientific and public health potential of biobanks through linkages to these diverse sources of biomedical information. A good example is the Norwegian Mother and Child Cohort Study (MoBa), which is the first of its kind to implement large-scale population screening of neurodevelopmental disorders early in life. It represents a unique resource for more in-depth analyses of gene-environment interaction in a temporal context. However, to fully exploit the scientific potential of the MoBa biobank and other national biorepositories, advanced analytical tools will need to be developed to mine the vast amounts of data generated through GWAS, whole-genome/exome-sequencing, and genome-wide epigenetic studies.

In this review, we described a suite of methods based on the Haplin software to investigate different causal scenarios that are relevant to perinatal disorders and other complex traits originating in early life. As the intrauterine environment depends on the mother's genetic make-up and her risk-taking behaviors, it is important to explore maternal and fetal gene-effects separately. The hybrid design is not only better-powered than the case-parent triad design, it also allows the main effect of an exposure to be assessed efficiently. The methods presented here have broad utility, ranging from studies of causes of disease to studies of interactions between medical treatments and patient genotypes in clinical studies. The accompanying tutorial highlights various applications of the software to analyze the types of data already available in MoBa and other national biobanks.

ACKNOWLEDGMENT

We are especially grateful to the many families who have participated in studies of clefting over the years. This research was supported through funds from Biohealth Norway (NFR 183255/O19); the European Union's Seventh Framework Programme (FP7/2007-2013), ENGAGE Consortium, grant agreement HEALTH-F4-2007-201413; the European Union's Seventh Framework Programme (FP7/2007-2013), BioSHaRE-EU, grant agreement HEALTH-F4-2010-261433; and the Norwegian Research Council (NFR 177522/V50).

REFERENCES

1. Stoltenberg C, Schjølberg S, Bresnahan M, et al. The Autism Birth Cohort: a paradigm for gene-environment-timing research. *Mol Psychiatry* 2010; **15** (7): 676-80.
2. Foley DL, Craig JM, Morley R, et al. Prospects for epigenetic epidemiology. *Am J Epidemiol* 2009; **169** (4): 389-400.
3. Rahimov F, Jugessur A, Murray JC. Genetics of nonsyndromic orofacial clefts. *Cleft Palate Craniofac J* 2012, http://dx.doi.org/10.1597/10-178.
4. Greene ND, Stanier P, Copp AJ. Genetics of human neural tube defects. *Hum Mol Genet* 2009; **18** (R2): R113-29.
5. Vefring HK, Wee L, Jugessur A, Gjessing HK, Nilsen ST, Lie RT. Maternal angiotensinogen (AGT) haplotypes, fetal renin (REN) haplotypes and risk of preeclampsia; estimation of gene-gene interaction from family-triad data. *BMC Med Genet* 2010; **11**: 90.
6. Weinberg CR, Shi M. The genetics of preterm birth: using what we know to design better association studies. *Am J Epidemiol* 2009; **170** (11): 1373-81.
7. 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**: 382-96.
8. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 2002; **11** (6): 513-20.
9. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev* 2002; **11** (6): 505-12.
10. Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000; **152** (3): 197-203.
11. 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.
12. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med* 1997; **16** (15): 1731-43.

13. 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.
14. Shen YC, Fan JH, Edenberg HJ, 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** (7): 1272-7.
15. Kazma R, Babron MC, Genin E. Genetic association and gene-environment interaction: a new method for overcoming the lack of exposure information in controls. *Am J Epidemiol* 2011; **173** (2): 225-35.
16. Vermeulen SH, Shi M, Weinberg CR, Umbach DM. A hybrid design: case-parent triads supplemented by control-mother dyads. *Genet Epidemiol* 2009; **33** (2): 136-44.
17. Weinberg CR, Shi M, Umbach DM. Re.: "Genetic association and gene-environment interaction: a new method for overcoming the lack of exposure information in controls". *Am J Epidemiol* 2011; **173** (11): 1346-7; author reply 7-8.
18. Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. *Am J Epidemiol* 2008; **168** (5): 541-7.
19. 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.
20. Weinberg CR. Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* 1999; **64** (4): 1186-93.
21. Weinberg CR, Morris RW. Testing for Hardy-Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. *Am J Epidemiol* 2003; **158** (5): 401-3; discussion 404-5.
22. 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.
23. 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.
24. R Development Core Team. R: A Language and Environment for Statistical Computing. 2006.
25. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100** (16): 9440-5.
26. Efron B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge: Cambridge University Press, 2010.
28. Wehby GL, Cassell CH. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis* 2010; **16** (1): 3-10.
29. Strauss RP. The organization and delivery of craniofacial health services: the state of the art. *Cleft Palate Craniofac J* 1999; **36** (3): 189-95.
30. Zhu JL, Basso O, Hasle H, Winther JF, Olsen JH, Olsen J. Do parents of children with congenital malformations have a higher cancer risk? A nationwide study in Denmark. *Br J Cancer* 2002; **87** (5): 524-8.
31. Bille C, Winther JF, Bautz A, Murray JC, Olsen J, Christensen K. Cancer risk in persons with oral cleft – a population-based study of 8,093 cases. *Am J Epidemiol* 2005; **161** (11): 1047-55.
32. Frebourg T, Oliveira C, Hochain P, et al. Cleft lip/palate and CDH1/E-cadherin mutations in families with hereditary diffuse gastric cancer. *J Med Genet* 2006; **43**: 138-42.
33. Christensen K, Juel K, Herskind AM, Murray JC. Long term follow up study of survival associated with cleft lip and palate at birth. *BMJ* 2004; **328** (7453): 1405.
34. Dietz A, Pedersen DA, Jacobsen R, Wehby GL, Murray JC, Christensen K. Risk of breast cancer in families with cleft lip and palate. *Ann Epidemiol* 2012; **22**: 37-42.
35. 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** (10): 2468-74.
36. Grosen D, Bille C, Petersen I, et al. Risk of oral clefts in twins. *Epidemiology* 2011; **22** (3): 313-9.
37. Grosen D, Chevrier C, Skytthe A, et al. A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. *J Med Genet* 2010; **47** (3): 162-8.
38. Sivertsen A, Wilcox AJ, Skjærven R, et al. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *BMJ* 2008; **336** (7641): 432-4.
39. Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. *Nature Rev Genet* 2011; **12** (3): 167-78
40. Wilcox AJ, Lie RT, Solvoll K, et al. Folic acid supplements and risk of facial clefts: national population based case-control study. *BMJ* 2007; **334** (7591): 464.
41. Jugessur A, Shi M, Gjessing HK, et al. Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PLoS One* 2009; **4** (4): e5385.

42. Beaty TH, Murray JC, Marazita ML, et al. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genet* 2010; **42** (6): 525-9.
43. Letterio JJ, Geiser AG, Kulkarni AB, Roche NS, Sporn MB, Roberts AB. Maternal rescue of transforming growth factor-beta 1 null mice. *Science* 1994; **264** (5167): 1936-8.
44. Popliker M, Shatz A, Avivi A, Ullrich A, Schlessinger J, Webb CG. Onset of endogenous synthesis of epidermal growth factor in neonatal mice. *Dev Biol* 1987; **119** (1): 38-44.
45. Buyske S. Maternal genotype effects can alias case genotype effects in case-control studies. *Eur J Hum Genet* 2008; **16** (7): 783-5.
46. 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.
47. 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.
48. Jugessur A, Shi M, Gjessing HK, et al. Maternal genes and facial clefts in offspring: a comprehensive search for genetic associations in two population-based cleft studies from Scandinavia. *PLoS One* 2010; **5** (7): e11493.
49. Jugessur A, Shi M, Gjessing HK, et al. Fetal genetic risk of isolated cleft lip only versus isolated cleft lip and palate: A subphenotype analysis using two population-based studies of orofacial clefts in Scandinavia. *Birth Defects Res A Clin Mol Teratol* 2011; **91**: 85-92.
50. Sinsheimer JS, Palmer CG, Woodward JA. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol* 2003; **24** (1): 1-13.
51. Starr JR, Hsu L, Schwartz SM. Assessing maternal genetic associations: a comparison of the log-linear approach to case-parent triad data and a case-control approach. *Epidemiology* 2005; **16** (3): 294-303.
52. Starr JR, Hsu L, Schwartz SM. Performance of the log-linear approach to case-parent triad data for assessing maternal genetic associations with offspring disease: type I error, power, and bias. *Am J Epidemiol* 2005; **161** (2): 196-204.
53. Shi M. Comprehensive gene environment analysis of the causes of orofacial clefts [Doctoral thesis]. Iowa City: Graduate College of the University of Iowa, 2005.
54. Horvath S, Laird NM, Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. *Am J Hum Genet* 2000; **66** (3): 1161-7.
55. Knapp M. Reconstructing parental genotypes when testing for linkage in the presence of association. *Theor Popul Biol* 2001; **60** (3): 141-8.
56. Ho GY, Bailey-Wilson JE. The transmission/disequilibrium test for linkage on the X chromosome. *Am J Hum Genet* 2000; **66** (3): 1158-60.
57. Ding J, Lin S, Liu Y. Monte Carlo pedigree disequilibrium test for markers on the X chromosome. *Am J Hum Genet* 2006; **79** (3): 567-73.
58. Chung RH, Morris RW, Zhang L, Li YJ, Martin ER. X-APL: an improved family-based test of association in the presence of linkage for the X chromosome. *Am J Hum Genet* 2007; **80** (1): 59-68.
59. Zhang L, Martin ER, Morris RW, Li YJ. Association test for X-linked QTL in family-based designs. *Am J Hum Genet* 2009; **84** (4): 431-44.
60. 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** (4): 370-80.
61. Carlier M, Roubertoux PL, Wahlsten D. Maternal effects in behavior genetic analysis. In: Jones BC, Mormede P, eds. Neurobehavioral Genetics: Methods and Applications. New York: CRC Press, 1999.
62. Barlow DP. Gametic imprinting in mammals. *Science* 1995; **270** (5242): 1610-3.
63. Marazita ML. Segregation analyses. In: Wyszynski DF, editor. Cleft Lip and Palate: From Origin to Treatment. New York: Oxford University Press, 2002: 222-33.
64. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273** (5281): 1516-7.
65. Sham PC, Cherny SS. Genetic Architecture of Complex Diseases. In: Zeggini E, Morris A, eds. Analysis of Complex Disease Association Studies – A Practical Guide. London: Academic Press, 2011: 1-13.
66. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358** (9290): 1356-60.
67. Le TH, Coffman TM. Targeting genes in the renin-angiotensin system. *Curr Opin Nephrol Hypertens* 2008; **17** (1): 57-63.
68. Boyles AL, Wilcox AJ, Taylor JA, et al. Oral facial clefts and gene polymorphisms in metabolism of folate/one-carbon and vitamin A: a pathway-wide association study. *Genet Epidemiol* 2009; **33** (3): 247-55.
69. Juriloff DM. Mapping studies in animal models. In: Wyszynski DF, ed. Cleft Lip and Palate: From Origin to Treatment. New York: Oxford University Press, 2002: 265-82.

70. Millicovsky G, Johnston MC. Maternal hyperoxia greatly reduces the incidence of phenytoin-induced cleft lip and palate in A/J mice. *Science* 1981; **212** (4495): 671-2.
71. Zhu H, Kartiko S, Finnell RH. Importance of gene-environment interactions in the etiology of selected birth defects. *Clin Genet* 2009; **75** (5): 409-23.
72. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nature Rev Genet* 2010; **11** (4): 259-72.
73. Weinberg CR. Less is more, except when less is less: Studying joint effects. *Genomics* 2009; **93** (1): 10-2.
74. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies - challenges and opportunities. *Am J Epidemiol* 2009; **169** (2): 227-30; discussion 34-5.
75. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. *Nature Rev Genet* 2006; **7** (5): 385-94.
76. Jugessur A, Lie RT, Wilcox AJ, et al. Cleft palate, transforming growth factor alpha gene variants, and maternal exposures: assessing gene-environment interactions in case-parent triads. *Genet Epidemiol* 2003; **25** (4): 367-74.
77. Shi M, Christensen K, Weinberg CR, et al. Orofacial cleft risk is increased with maternal smoking and specific detoxification-gene variants. *Am J Hum Genet* 2007; **80** (1): 76-90.
78. Wu T, Liang KY, Hetmanski JB, et al. Evidence of gene-environment interaction for the IRF6 gene and maternal multivitamin supplementation in controlling the risk of cleft lip with/without cleft palate. *Hum Genet* 2010; **128** (4): 401-10.
79. Shi M, Umbach DM, Weinberg CR. Testing haplotype-environment interactions using case-parent triads. *Hum Hered* 2010; **70** (1): 23-33.
80. Shi M, Umbach DM, Weinberg CR. Family-based gene-by-environment interaction studies: revelations and remedies. *Epidemiology* 2011; **22** (3): 400-7.
81. Kistner EO, Shi M, Weinberg CR. Using cases and parents to study multiplicative gene-by-environment interaction. *Am J Epidemiol* 2009; **170** (3): 393-400.
82. Christensen K, Murray JC. What genome-wide association studies can do for medicine. *N Engl J Med* 2007; **356** (11): 1094-7.
83. Hirschhorn JN. Genomewide association studies – illuminating biologic pathways. *N Engl J Med* 2009; **360** (17): 1699-701.
84. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010; **8** (1): e1000294.
85. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461** (7265): 747-53.
86. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nature Rev Genet* 2005; **6** (2): 109-18.
87. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008; **299** (11): 1335-44.
88. Manolio TA, Rodriguez LL, Brooks L, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genet* 2007; **39** (9): 1045-51.
89. Birnbaum S, Ludwig KU, Reutter H, et al. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature Genet* 2009; **41** (4): 473-7.
90. Grant SF, Wang K, Zhang H, et al. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr* 2009; **155** (6): 909-13.
91. Mangold E, Ludwig KU, Birnbaum S, et al. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature Genet* 2010; **42** (1): 24-6.
92. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007; **23** (10): 1294-6.