Polyunphased : an extension to polytomous outcomes of the Unphased package for family-based genetic association analysis

Alexandre Bureau^{1,2,*} and Jordie Croteau¹

¹Institut universitaire en santé mentale de Québec du Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale, 2601 chemin de la Canardière, Québec, QC, G1J 2G3, Canada

²Département de médecine sociale et préventive, Université Laval, Pavillon Vandry, room 2457, 1050 rue de la Médecine, Québec, QC, G1V 0A6, Canada

Abstract

Polytomous phenotypes arise when a disease has multiple subtypes or when two dichotomous phenotypes are analyzed simultaneously. Few software programs offer the option to analyze such phenotypes in family studies, and none implements conditional polytomous logistic regression for within-family analysis robust to population stratification. We introduce *Polyunphased*, an extension to polytomous phenotypes of the *Unphased* package, a flexible software tool for genetic association analysis in nuclear families. Like *Unphased*, *Polyunphased* is written in C++ and runs from the command line or from a Java graphical user interface. Most *Unphased* options remain available in *Polyunphased*, including those handling missing parental genotypes while preserving robustness to population stratification, and the modelling options. Simulation studies confirmed the expected statistical behaviour of the maximum likelihood estimates of the association parameters of the conditional logistic regression model when the corresponding association parameters in the parental term of the likelihood function are set to 0, but revealed convergence problems when estimating these parental association parameters separately. The former approach is thus recommended with polytomous phenotypes.

Keywords

conditional likelihood; nuclear families; polytomous logistic model

Introduction

Many software packages perform genetic association analyses with dichotomous phenotypes in family samples, but few implement analysis of polytomous phenotypes. Such phenotypes can arise when a disease has multiple subtypes (Guey et al., 2010, Cocco et al., 2016) or when two dichotomous phenotypes are considered simultaneously, creating a four category phenotype (Bureau and Croteau, 2016). We have previously developed within-family score tests under polytomous models of association with one or two genetic variants (Bureau et al.,

*Corresponding author: Alexandre.Bureau@msp.ulaval.ca.

2014). Inference on the effect sizes of a variant on the various phenotypic categories via estimation of regression coefficients can be performed using statistical procedures for generalized mixed models such as `proc glimmix` in SAS or for marginal models fitted by generalized estimating equations such as functions from the R library `multgee`. However, these approaches do not protect against confounding due to population stratification.

Here we extend to polytomous phenotypes the approach for family-based association analysis in nuclear families proposed by Dudbridge (Dudbridge, 2008) and implemented in the `Unphased` package. Dudbridge's formulation handles multiple siblings per nuclear family, and allows testing for association in the presence of genetic linkage by conditioning on inheritance vectors which specify the grand-parental origin of each allele of each sibling, an essential feature in the context of follow-up of linkage signals in the same family sample. The approach allows missing parental genotype data, partially missing genotypes and missing phase information in haplotypes by modeling the genotype distribution of the parents, but achieves robustness to population stratification by letting the association parameters in the parental part of the likelihood function be distinct from the association parameters of interest in the conditional logistic regression model of the child genotypes given the parental genotypes and the child phenotypes. Here, we present `Polyunphased`, an extension of `Unphased` to polytomous phenotypes.

Implementation

`Polyunphased` was developed by extending the C++ and Java source code of `Unphased` to handle polytomous phenotypes at each step of the analysis. A polytomous phenotype can be specified using either the option “Polytomous trait” or the option “Joint traits” in the “Trait” menu of the Java graphical user interface (GUI), or by the corresponding two options on the command line: “-polytomous” followed by the name of a single polytomous outcome variable (taking values 0 to K , with 0 representing a missing value) or “-joint” followed by the name of two binary outcome variables separated by a column, for instance “-joint t1:t2”. The upper limit of K is a constant, which can be modified in the code. The default value is four, as fitting models with more levels is difficult. Outcome variables must be contained in a pedigree file in QTDT format (csg.sph.umich.edu/abecasis/QTDT/docs/input.html) and be specified in a data file to be of type “A” for affection status (even though with the option “-polytomous” the outcome variable need not be dichotomous as an usual affection status). Marker genotypes can be included in the pedigree file, or be read from a PLINK binary bedfile (pngu.mgh.harvard.edu/~purcell/plink/data.shtml#bed) specified with the “-bedfile” option as in `Unphased`.

With the “-polytomous” option, the largest integer value of the trait T is taken as the reference category K , and indicator response variables are defined as $Y_k = I(T=k)$ $k, = 1, \dots, K - 1$, where $I(\cdot)$ denotes the indicator function. With the “-joint” option, indicator variables are created from traits T_1 and T_2 as follows:

$$\begin{aligned} Y_1 &= I(T_1=2, T_2=1) \\ Y_2 &= I(T_1=1, T_2=2) \\ Y_3 &= I(T_1=2, T_2=2) \end{aligned}$$

The polytomous logistic model is then specified by $K - 1$ logistic regression functions, one for each Y_k , with distinct parameter vectors β_k , $k=1, \dots, K - 1$. Although the model is logistic, coefficient estimates do not tend to a log odds ratio, but rather to a quantity close to the log relative risk of phenotypic category k versus K for the corresponding genotypes or haplotypes, a behaviour resulting from the assumption that pseudo siblings with non-transmitted genotypes would have been in the reference category (Witte et al., 1999). (The difference is only noticeable with common phenotypic categories. In the remainder of this paper, we will refer to odds ratios in keeping with the `Unphased` terminology.) Ignoring genotype by covariate interactions, the probabilities of the child genotypes $C = (C_1, \dots, C_A)$ of the A children of a nuclear family conditional on paternal genotype F , maternal genotype M and child phenotypes (conditional term of the likelihood, equation 4 of (Dudbridge, 2008)) can be written for a $K=3$ level phenotype as

$$P_{\beta} [C=c|f, m, y_1, y_2] = \frac{\exp [y'_1 X'_c \beta_1 + y'_2 X'_c \beta_2]}{\sum_{c^* \in S(f, m)} \exp [y'_1 X'_{c^*} \beta_1 + y'_2 X'_{c^*} \beta_2]} \quad (1)$$

where $X_c = (X_{c1}, \dots, X_{cA})$ and X_{c_a} is a vector of numerical codings for genotype c_a (for instance, X_{c_a} could be the number of copies of each haplotype of c , of length $H-1$ in presence of H haplotypes), and $\beta = (\beta'_1, \dots, \beta'_{K-1})'$ is the full vector of regression parameters, which would be of length $(K-1)(H-1)$.

Similarly, the parental genotype probabilities, conditional on child phenotype (parental term of the likelihood, equation 5 of (Dudbridge, 2008)) can be written as

$$P_{\lambda, \beta} [F=f, M=m|y_1, y_2] = \frac{\sum_{c^* \in S(f, m)} \exp [X'_{fm} \lambda + y'_1 X'_{c^*} \tilde{\beta}_1 + y'_2 X'_{c^*} \tilde{\beta}_2]}{\sum_{f^*, m^*} \sum_{c^* \in S(f, m)} \exp [X'_{f^* m^*} \lambda + y'_1 X'_{c^*} \tilde{\beta}_1 + y'_2 X'_{c^*} \tilde{\beta}_2]} \quad (2)$$

where X_{fm} denotes a vector of codes for the mating types (f, m) and λ represents the distribution of mating types in the parents of subjects in the reference category K . It is parameterized as a function of haplotype frequencies under the assumptions of random mating and Hardy-Weinberg equilibrium in `Unphased` and `Polyunphased`. The parameters in the parental term are denoted $\tilde{\beta}$ to indicate that they can be estimated separately from β when the “-parentrisk” option is specified or set to 0 without that option. Estimating β from the conditional term of the likelihood alone protects against misspecification of the mating type distribution, which occurs in situations of population stratification (see (Dudbridge, 2008) for further explanations).

The analytical gradient of the log-likelihood from the terms (1) and (2) is coded in `Polyunphased` and used in the maximization of the likelihood by the Davidon-Fletcher-Powell method. Alternatively, the Nelder-Mead's downhill simplex optimization method, which does not require the gradient, can be selected with the “-neldermead” option. The estimated covariance matrix of the coefficient estimates is computed numerically from the gradient and used to obtain confidence intervals for β .

Experimentation with the program revealed that the maximum of the log-likelihood under the specified model was often lower than the maximum of the log-likelihood under the null model with a single set of starting values, especially using the “-parentrisk” option, suggesting the existence of local maxima of the log-likelihood in high dimensions. We conducted a simulation study described in the appendix to investigate the factors influencing log-likelihood maximization. When estimating only the β vector in the conditional term of the likelihood and setting $\tilde{\beta}$ to 0, the maximum of the log-likelihood was higher under the specified model than under the null model in all 100 replicates with a single set of starting values. Results in Table 1 indicate that the bias of the coefficient estimates was minimal compared to their variance. Coverage of the true log relative risk by confidence intervals was close to the nominal 95% level. By contrast, when fitting separate β and $\tilde{\beta}$ for the conditional and parental parts of the likelihood with the “-parentrisk” option, 500 random sets of starting values (using the “-restarts” option) were needed to obtain a maximal log-likelihood higher under the specified model than under the null model in 72% of replicates with 100 families, and in 99% of replicates with 1000 families. The Nelder-Mead optimization method with the same starting values reached the same solution, with increased computing time. Even when the likelihood ratio statistic was positive, it is likely that the solution was a local maximum in a large number of replicates, given the remaining important bias and poor confidence interval coverage with 1000 families.

Usage

We illustrate the use of `Polyunphased` by reanalyzing the association of the four-level phenotype formed by all combinations of major psychosis and visual episodic memory dichotomous phenotypes with single nucleotide polymorphisms (SNPs) from the 13q13-q14 genomic region in 14 informative extended families from the Eastern Quebec kindred study (Bureau et al., 2014). The major psychosis phenotype is called MP and the visual episodic memory phenotype is called `reyrdcut4`. Given the poor behaviour of the estimates when fitting separate β and $\tilde{\beta}$ (“-parentrisk” option) in the simulations, and given Dudbridge (2008) found only slight Type I error inflation in families without genotyped parents from a stratified population when setting $\tilde{\beta}$ to 0, we did not use the “-parentrisk” option. The analysis of the SNP rs1156026 was run with the following command

```
> polyunphased -datafile datafile.txt -pedfile chr13.pre -marker
rs1156026 -joint MP:reyrdcut4 -missing -restarts 20 -randomseed 777
```

The same analysis can be conducted by opening the pedigree file `chr13.pre` and the data file `datafile.txt` through the “File” menu of the GUI, next clicking the “Joint traits” option in the

“Trait” menu and selecting traits MP and reyrdcut4 from the displayed trait list, clicking “Test marker” in the “Marker” menu and selecting rs1156026. The additional options “Uncertain haplotypes and missing genotypes”, “Random restarts” and “Random number seed” can be set under the “Option” menu, as in Unphased.

Running time was 35 seconds using a single 2.8 GHz Intel E7-4870 processor of a HP DL580R07 computer. Selected output is presented in Figure 1. The default behaviour of Unphased is to split extended families into nuclear families, and this has been preserved in Polyunphased, the output informing us that 171 nuclear families were obtained. Allele frequencies are presented separately for the alleles transmitted and untransmitted from the parents to their children in the levels 1 to 4 representing all combinations of the two dichotomous outcomes. The likelihood ratio test of overall association is the three-degree-of-freedom test of the three model coefficients. Below, odds ratios for the T allele (numerically coded as 4) against the reference C allele (numerically coded as 2) are presented for levels 1 to 3 against the reference level 4, corresponding to ($T_1=1, T_2=1$). The 95% confidence intervals cover 1, consistent with the p-values greater than 0.05 in Table 3 of (Bureau et al., 2014) obtained with the within-family score tests of the coefficients of the single marker polytomous model.

We also analyzed association to rs2120753, a SNP moderately correlated with rs1156026 (squared correlation = 0.30) which had shown association to schizophrenia in univariate analyses in another sample (Bureau et al., 2013). This analysis is performed conditional on the alleles observed at rs1156026, to determine presence of association between rs2120753 and the phenotypic categories independent from or in interaction with rs1156026, using the following command:

```
> polyunphased -datafile datafile.txt -pedfile chr13.pre -marker
rs2120753 -condition rs1156026 -joint MP:reyrdcut4 -full -missing
-restarts 20 -randomseed 777
```

In the GUI, this analysis involves setting the “Test markers” to rs2120753 and the “Conditioning markers” to rs1156026 in the “Marker” menu and setting the “Full model” option in the “Analysis” menu, in addition to selecting the same other options as before. Running time was 2.9 minutes. Selected output is presented in Figure 2. The test requested by the option “-full” is a six degree of freedom test of the rs2120753 three main effects and rs2120753 by rs1156026 three product terms. As with Unphased, default analyses are based on haplotype frequencies, and odds ratios of each level against the reference level 4 for each haplotype are presented against a reference haplotype, which can be specified with the “-reference” option (here the reference haplotype is 2-2 by default).

Discussion

We have developed Polyunphased from the Unphased package for analyzing genetic association to polytomous outcomes in families, the first package fitting polytomous conditional logistic regression models adapted to nuclear-family designs. Extended families

can be split into nuclear families for analysis; we showed elsewhere in the dichotomous case that the dependence between nuclear families from the same extended family introduces negligible bias in the variance estimates (Bureau and Duchesne, 2015), leading to valid Wald tests and confidence intervals. Although we focus on analysis of families where at least some parents are genotyped, the modification of the likelihood for samples of sibships without any parent genotype (Dudbridge et al., 2011) requested by the “-sibship” option remains available with polytomous phenotypes. `Polyunphased` inherits most analysis options from `Unphased`, conferring it a great flexibility for the analysis of multiple markers in a genomic region, for instance analyzing association to a marker conditional on another one as illustrated in the Usage section, adjusting for covariates or analyzing the modifying effect of covariates. Such within-region analysis, with estimated frequencies of the involved markers’ haplotypes and assumption of tight linkage between the markers in the conditional likelihood, is not equivalent to the analysis of a marker conditional on another independent and unlinked marker using score tests proposed previously (Bureau et al., 2014).

The number of regression coefficients to estimate is proportional to the number of levels of the polytomous outcome minus one. This increase in dimension compared to a dichotomous outcome may lead to convergence problems. This is especially true for the estimation of separate coefficients in the conditional and parental terms of the likelihood, an option that confers better robustness to population stratification than setting the coefficients of the parental terms to 0, but doubles the number of regression coefficients. Given the bias and poor confidence interval coverage observed in our simulation study, even in samples as large as 1000 families, we must advise against using the “-parentrisk” option with a polytomous phenotype. We note that if the parental genotypes are fully observed, the likelihood factors into the conditional and parental terms, and estimates of the regression coefficients of the conditional term are independent of the value of the regression coefficients of the parental term. In that case, the estimates with the “-parentrisk” option would be identical to the estimates without the “-parentrisk” option if the likelihood maximization were to converge. If one attempts nonetheless to use this option, we recommend maximizing the likelihood from a large number of starting sets of parameter values to improve the chances of finding a globally optimal solution. This recommendation also applies to analyses involving covariates, as this too leads to convergence problems (even without the “-parentrisk” option) due to the larger number of coefficients.

Not all `Unphased` features are yet available in `Polyunphased`. In particular, only the nuclear family analysis is implemented. We plan to adapt the likelihood contribution of unrelated subjects to polytomous phenotypes, enabling the analysis on a combined sample of nuclear families and unrelated subjects in the near future.

Availability

The C++ source code, executable files for Ubuntu Linux and Windows, and the Java GUI are available under the GNU General Public License at github.com/abureau/polyUnphased.

Acknowledgments

This work has been supported by the Canadian Institutes of Health research (CIHR, grant MOP-106448). We thank M. Maziade (U Laval) for access to the data from the Eastern Quebec Schizophrenia and Bipolar Disorder Study, funded by CIHR (grants MOP-74430, MOP-119408 and MOP-114988) and by a Canada Research Chair (950-200810) in the genetics of neuropsychiatric disorders of which M. Maziade is the Chair.

References

- Bureau A, et al. Follow-up of a Major Psychosis Linkage Site in 13q13-q14 Reveals Significant Association in Both Case-Control and Family Samples. *Biol Psychiatry*. 2013; 74(6):444–450. [PubMed: 23602252]
- Bureau A, Croteau J. When Is an Endophenotype Useful to Detect Association to a Disease? Exploring the Relationships between Disease Status, Endophenotype and Genetic Polymorphisms. *Hum Hered*. 2016; 81(1):11–25. [PubMed: 27475094]
- Bureau A, et al. Extension of the generalized disequilibrium test to polytomous phenotypes and two-locus models. *Front Genet*. 2014; 5:258. [PubMed: 25152751]
- Bureau A, Duchesne T. On the validity of within-nuclear-family genetic association analysis in samples of extended families. *Stat Appl Genet Mol Biol*. 2015; 14(6):533–549. [PubMed: 26544107]
- Cocco P, et al. N-acetyltransferase polymorphisms are associated with risk of lymphoma subtypes. *Hematol Oncol*. 2016; 34(2):79–83. [PubMed: 25689677]
- Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered*. 2008; 66(2):87–98. [PubMed: 18382088]
- Dudbridge F, et al. A flexible model for association analysis in sibships with missing genotype data. *Ann Hum Genet*. 2011; 75(3):428–438. [PubMed: 21241274]
- Guey LT, et al. Genetic susceptibility to distinct bladder cancer subphenotypes. *Eur Urol*. 2010; 57(2):283–292. [PubMed: 19692168]
- Witte JS, et al. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol*. 1999; 149(8):693–705. [PubMed: 10206618]

Appendix

Simulation study

In nuclear families with four children each, we simulated a genetic variant with two alleles assuming the genotype is observed in all family members. The allele with frequency 0.3 increased the risk of having outcome 1 without outcome 2 ($T_1 = 2, T_2 = 1$) or of having both outcomes ($T_1 = 2, T_2 = 2$) to different extents (see table 1 where β_k , $k=1,2,3$, represent log relative risks between categories k and 4). Outcomes 1 and 2 had population prevalence of 0.30 and 0.12 respectively. We formed samples of nuclear families ascertained for having at least one child belonging to a phenotypic category other than the reference. We performed two simulations in which we formed samples of 100 and 1000 nuclear families. In both cases, the process was repeated 100 times. We used the option “-joint” to form a polytomous phenotype with the four possible combinations of T_1 and T_2 .

Reading data file datafile.txt (QTDT format)...done
 Reading pedigree file chr13.pre...done 2 2

ANALYSIS OPTIONS

Number of nuclear families:	171
Number of unrelated subjects:	0
Estimate missing genotypes (-missing):	Y
Numerical maximisation (-neldermead):	D-F-P
Random restarts (-restarts):	20
Random number seed (-randomseed):	777

Exploratory pass...done
 Identifying rare haplotypes: Alternative likelihood...done (490 evaluations)
 Null likelihood...done (243 evaluations)
 Alternative likelihood...done (500 evaluations)

ANALYSIS OF POLYTOMOUS TRAIT: reyrd_cut4:MP

MARKER OPTIONS

Test marker: rs1156026

ESTIMATES OF MARGINAL ALLELE FREQUENCIES IN FAMILIES

	Level 1	Level 2	Level 3	Level 4	Level 1	Level 2	Level 3	Level 4
Allele	Trans	Trans	Trans	Trans	Freq	Freq	Freq	Freq
2	31	36	20	145	0.5741	0.6	0.5	0.6532
4	23	24	20	77	0.4259	0.4	0.5	0.3468

Allele	Untrans	Untrans	Untrans	Untrans	Freq	Freq	Freq	Freq
2	36.32	33.06	22.75	135.9	0.6727	0.551	0.5686	0.6121
4	17.68	26.94	17.25	86.11	0.3273	0.449	0.4314	0.3879

TEST OF OVERALL ASSOCIATION

Null log-likelihood = -262.832
 Alternative log-likelihood = -262.25
 Likelihood ratio chisq = 1.16501 df = 3 p-value = 0.761409

ESTIMATES OF ALLELE EFFECTS

Reference allele: 2

	Level 1			Level 2			Level 3		
Allele	Odds-R	95%Lo	95%Hi	Odds-R	95%Lo	95%Hi	Odds-R	95%Lo	95%Hi
2	1	1	1	1	1	1	1	1	1
4	1.332	0.6286	2.824	0.8308	0.35	1.972	1.289	0.4619	3.598

Figure 1.


```

Reading data file datafile.txt (QTD format)...done
Reading pedigree file chr13.pre...done 2 2

ANALYSIS OPTIONS
Number of nuclear families:          171
Number of unrelated subjects:        0
Estimate missing genotypes (-missing): Y
Numerical maximisation (-neldermead): D-F-P
Random restarts (-restarts):         20
Random number seed (-randomseed):    777

-----
Exploratory pass...done
Identifying rare haplotypes: Alternative likelihood...done (1719 evaluations)
Null likelihood...done (758 evaluations)
Alternative likelihood...done (1715 evaluations)

ANALYSIS OF POLYTOMOUS TRAIT: reyrd_cut4:MP

MARKER OPTIONS
Test marker: rs2120753
Conditioning marker: rs1156026

ESTIMATES OF MARGINAL HAPLOTYPE FREQUENCIES IN FAMILIES
      Level 1  Level 2  Level 3  Level 4  Level 1  Level 2  Level 3  Level 4
Haplotype Trans  Trans  Trans  Trans  Freq  Freq  Freq  Freq
2-2      24.82  26.98  15.95  101.2  0.4279  0.4496  0.3987  0.4363
2-4      9.726  9.022  4.05   51.63  0.1677  0.1504  0.1013  0.2226
4-2      4.184  0.02173  1.05   5.786  0.07214  0.00036  0.02626  0.02494
4-4      19.27  23.98  18.95  73.37  0.3323  0.3996  0.4737  0.3162

Haplotype Untrans  Untrans  Untrans  Untrans  Freq  Freq  Freq  Freq
2-2      25.9   20.64  14.18  97.27  0.4465  0.344  0.3545  0.4193
2-4      12.82  12.76  8.692  46.47  0.221  0.2126  0.2173  0.2003
4-2      1.913  3.555  2.387  7.474  0.03298  0.05925  0.05968  0.03221
4-4      17.37  23.04  14.74  80.79  0.2995  0.3841  0.3686  0.3482

TEST OF OVERALL ASSOCIATION
Null log-likelihood = -466.395
Alternative log-likelihood = -462.957
Likelihood ratio chisq = 6.87575 df = 6 p-value = 0.332491

ESTIMATES OF HAPLOTYPE EFFECTS
Reference haplotype: 2-2
      Level 1
Haplotype Odds-R  95%Lo  95%Hi
2-2      1      1      1
2-4      0.7236  0.1768  2.961
4-2      2.114  0.2057  21.72
4-4      0.9234  0.3125  2.729

      Level 2
Haplotype Odds-R  95%Lo  95%Hi
2-2      1      1      1
2-4      0.1577  2.124  0.4846
4-2      0      0      0.4703
4-4      0.3305  2.421  1.174

      Level 3
Haplotype Odds-R  95%Lo  95%Hi
2-2      1      1      1
2-4      0.1154  2.035  0.03799
4-2      0.03799  5.823  0.03799
4-4      0.373  3.695  0.03799

```

Figure 2.

Table 1

Statistical properties of the polytomous logistic regression coefficient maximum likelihood estimates.

		β_1		β_2		β_3	
		Fixed ^a	Sep ^b	Fixed ^a	Sep ^b	Fixed ^a	Sep ^b
100 families per replicate	Value	0.693	0.693	0	0	0.549	0.549
	Bias	0.078	0.043	-0.098	0.006	0.070	0.032
	Variance of $\hat{\beta}$	0.055	0.139	0.228	0.668	0.247	0.484
	Mean Squared Error	0.060	0.139	0.235	0.659	0.250	0.478
	95% confidence interval coverage	0.96	0.833	0.99	0.903	0.91	0.847
1000 families per replicate	Value	0.693	0.693	0	0	0.549	0.549
	Bias	0.010	0.149	0.020	0.283	0.017	-0.008
	Variance of $\hat{\beta}$	0.005	0.251	0.018	0.339	0.014	0.306
	Mean Squared Error	0.005	0.271	0.018	0.415	0.014	0.303
	95% confidence interval coverage	0.98	0.333	0.98	0.384	0.95	0.374

^aFixed coefficients for parental part of the likelihood (set to 0).

Based on 100 replicates.

^bCoefficients for parental and conditional parts of the likelihood estimated separately. `restarts` option set to 500 for each replicate.

Results based on the 72 and 99 replicates with positive value of LR test for 100 and 1000 families per replicate respectively.