Approaches to binary trait association analysis using family data

David Duffy

Genetic Epidemiology Laboratory, QIMR

Introduction

- The setting
- A computer package for statistical genetics: Sib-pair
- Simulation-based percentiles for case-control allele frequency differences
- Quasi-likelihood tests of case-control allele frequency differences
- (Generalized) Estimating Equation models for association
- Generalized Linear Mixed Models for association
- Within-family tests of association and linkage
- Identity-By-State allele sharing tests for association

Motivating example: Brisbane Twin Nevus Study

- 1200 families: parents, 12 y.o. twin offspring plus 0-6 additional siblings
- In 450 families, twins are monozygotic (MZ, genetically identical)
- Extensive measures of quantitative and categorical traits
- Particular interest in eye, skin, hair colour, freckling
- 4300 individuals genotyped at varying numbers of genetic markers: up to 610K SNPs

Motivating example: Queensland Familial Melanoma Study

- 2700 melanoma cases in 1700 families: up to 63 members
- Population-based ascertainment, oversampling dense pedigrees
- Similar pigmentation phenotypes measured
- 2250 individuals genotyped: mainly known pigmentation genes
- We use BTNS families as controls

Motivating example: BTNS + QFMP

- Both cases and controls come in families
- Lots of missing genotype data: not MCAR
- Do have strong model (Mendelism) to allow imputation of genotypes
- Differing age structures: irrelevant to genotype frequencies?
- Differing ethnic origins: large effects on pigmentation genes
- Large number of MZ twins among controls (a few cases too!)
- Simpler family structures (mostly two generation)
- Detected gene effect sizes are small, so need maximum power

Motivating example: Atopy in West Highland White Terriers

- Kindred of affected and unaffected animals (380 animals, 13 generations)
- A collection of additional affected animals
- Genotypes at 50000 markers available for 19 unaffected and 32 affected
- Inbred (unaffected animals F=0.022, affected animals F=0.008)
- Setup more congenial to genetic linkage analysis, but association quick

Overview of Sib-pair

An extensible platform for genetic data manipulation and analysis

A platform for methodological experimentation

First code written in 1995. Now all standard Fortran 95, compiles using multiple compilers on multiple platforms.

Creeping featurism has continued to today (63000 lines of code; 18000 LOC in last 12 months)

Home page: **`http://www.qimr.edu.au/davidD#sib-pair`**

Three approaches to missing genotype data in Sib-pair

Imputing missing genotypes:

- Maximum likelihood (peeling)
- Markov Chain Monte Carlo simulation
- Quasi-likelihood BLUP

Binary trait association analysis incorporating missing genotypes:

- Multiple imputation (Rubin 1987)
- ML using genotype probabilities or expected allelic scores
- Quasi-likelihood MQLS (Thornton and McPeek 2007)

Multiple imputation for missing genotypes

- Popular approach for complex datasets in nongenetic contexts (Rubin 1987)
- Impute missing data probabilistically so generate multiple (different) versions of data (usually only 5-10)
- Calculate a test statistic (*S*) and error variance (*U*) for each dataset as if it was fully observed
- Estimate between-replicate and within-replicate variances
- Calculate corrected Wald test

$$\overline{S} = m^{-1}\Sigma S^{(i)}$$

$$V(\overline{S}) = (1 + m^{-1})[(m-1)^{-1}\Sigma(S^{(i)} - S)^{2}] + m^{-1}\Sigma U^{(i)}$$

Gene-dropping: simulating the founders

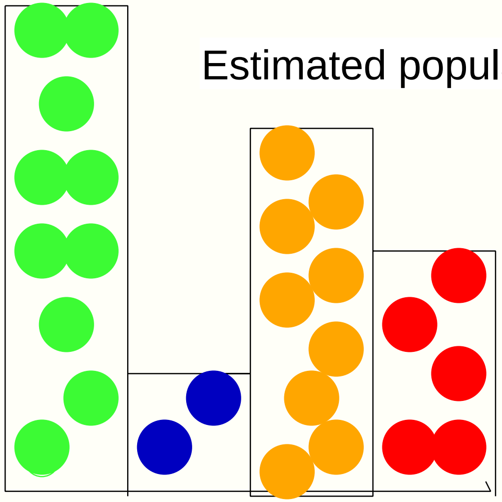
Gene-dropping is a method used to simulate a codominant marker in a family.

Pedigree founder genotypes are first generated by multinomial sampling from the measured population genotype frequencies.

Assuming Hardy-Weinberg Equilibrium, genotype frequencies can be calculated from allele frequencies:

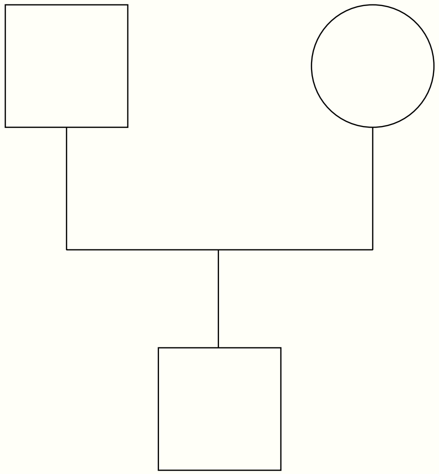
So we draw two alleles for each person, using the allele frequencies as the probability of choosing each type of allele.

Gene dropping 1

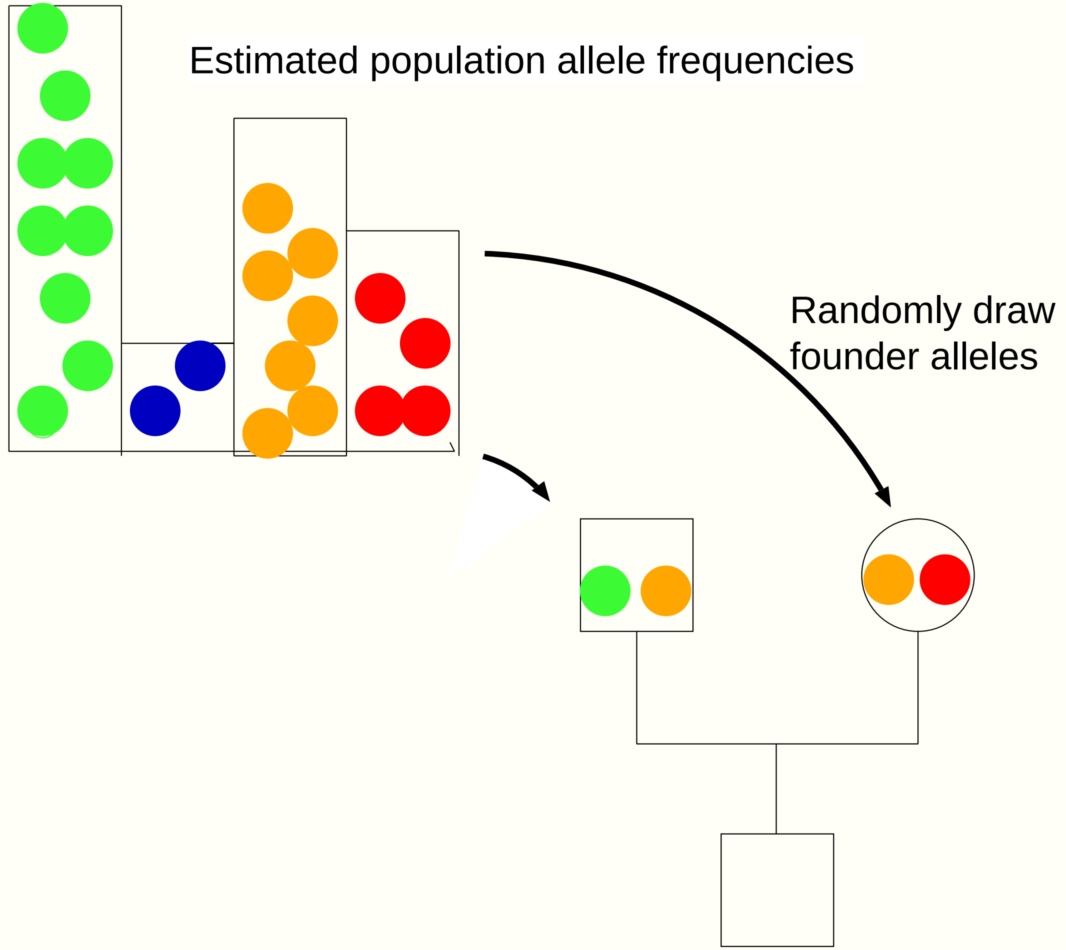


Estimated population allele frequencies

Pedigree for simulation



Gene dropping 2



Estimated population allele frequencies

Randomly draw founder alleles

Gene-dropping: simulating the nonfounders

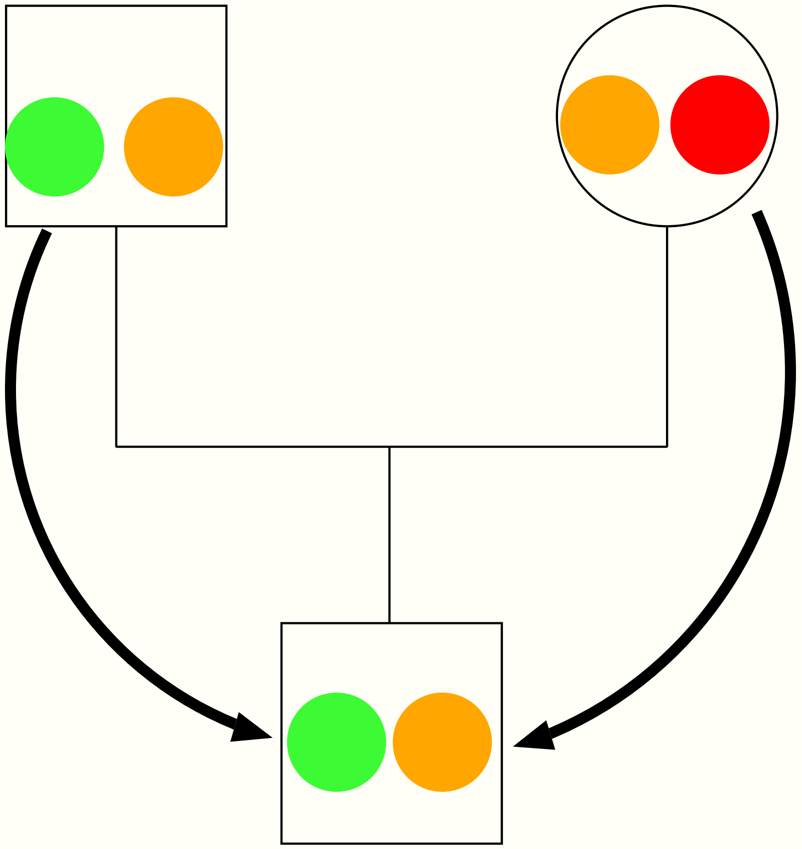
We simulate childrens’ genotypes by randomly drawing one allele from each parental genotype (they are equally likely).

And simulate childrens’ childrens’ genotypes by same process…

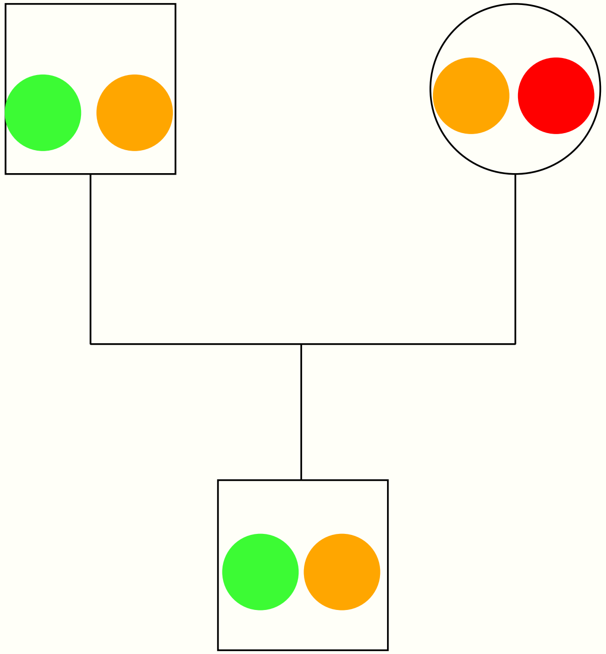
Until the pedigree genotypes are completely filled in.

A monozygotic twin always receives the same genotype as his twin.

Randomly draw one allele from each parent



Gene dropping 4

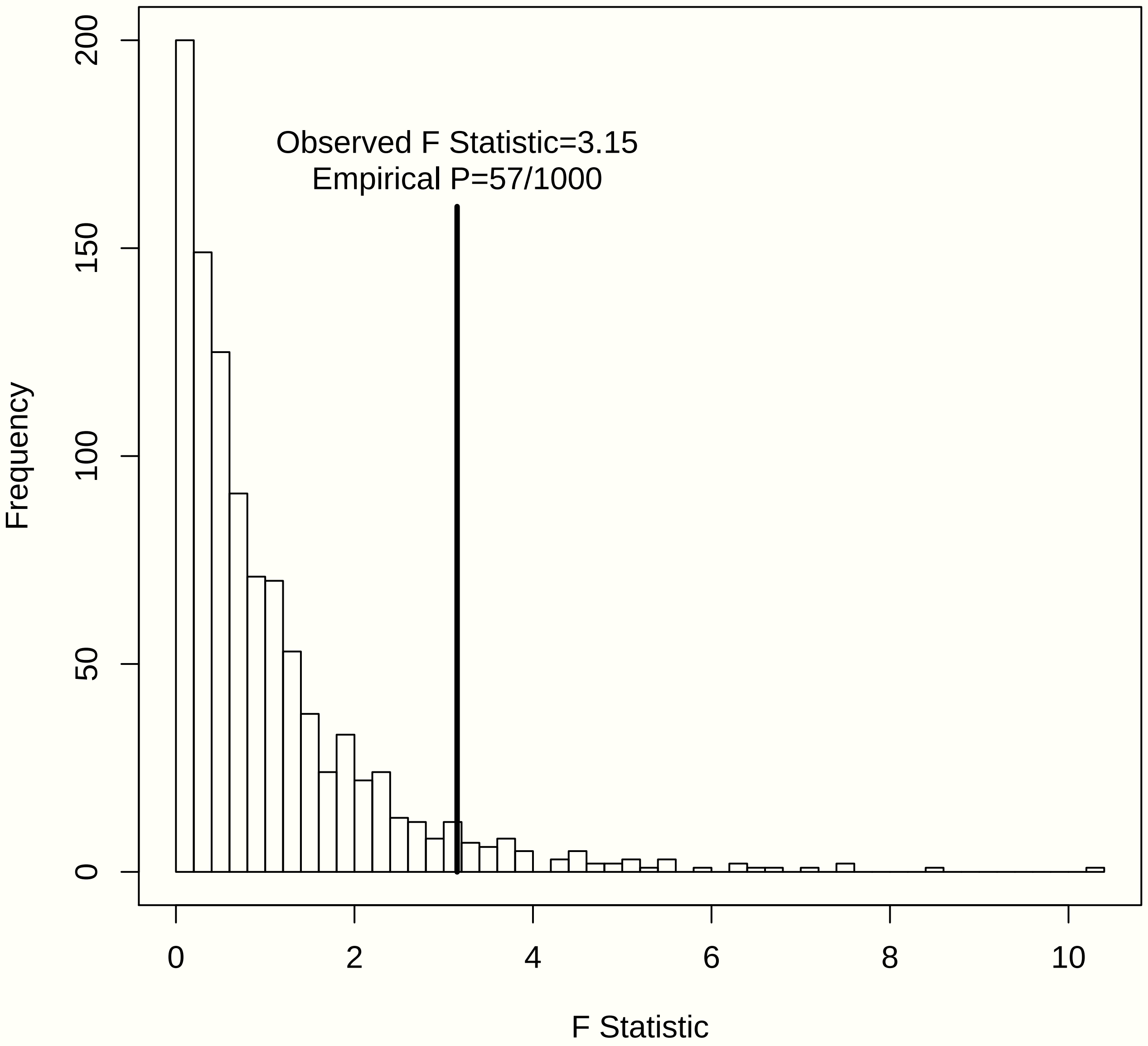


Y=14.5

Calculate test statistic given the simulated genotypes

$$Y = \beta_G G_i + \beta_{PG} PG_i$$

1000 simulations under null hypothesis of no association



Gene-Dropping: binary trait association analysis

For example, testing association between a binary trait and a codominant marker, correctly allowing for the pedigree structure of the data:

Test Statistic: Ordinary contingency table chi-square test, X^2_{Obs}

Problem: Usual reference distribution assumes independence of observations

Solution: Generate correct reference distribution by simulation

Gene-Dropping: refinements

- The distribution of the simulated test statistic is conditional on ths observed trait values, so the effects of *unmeasured* genes are included in the simulation.
- The alternative approach of permuting phenotypes must be careful to retain the phenotypic correlation structure eg permute families.
- To produce within-family tests, the simulation can skip step Ia. above, so the reference distribution is "Conditional on Parental Genotypes"
- B does not have to be fixed, so that the simulation stops when the P-value is sufficiently accurate (sequential approach of Besag and Clifford 1991).

Choice of trait-genotype association measure

Geneticists have generally used an allelic test. For a diallelic marker, genotypes AA, AB, BB are coded as 0, 1, 2. For such a marker, the test statistic usually reduces to the form:

$$\frac{(p_{Case} - p_{Pop})^2}{V}$$

This has been criticized, but takes advantage of the assumption of Hardy-Weinberg equilibrium to smooth out the effects of stochastic variation in genotype counts, even in the diallelic case.

One alternative is Empirical Bayes smoothing of the control genotype frequencies alone (Cheng and Chen 2005). This can be easily extended to the case of analysis stratifying on covariates.

Breast cancer and BRCA1

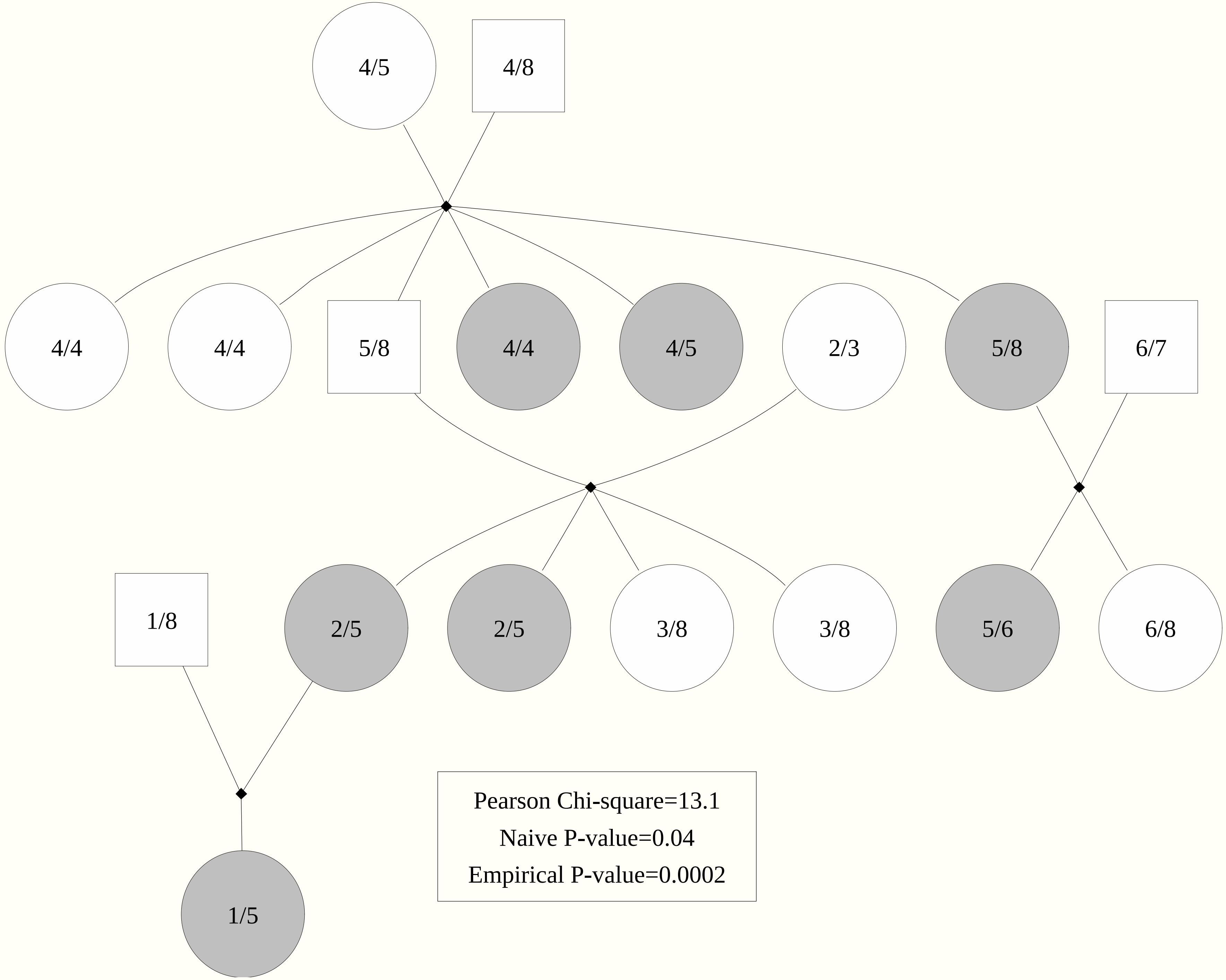
Hall et al (1990) reported that breast cancer in densely affected pedigrees was linked to a marker (D17S74) on chromosome 17. In the first pedigree they described, the P-value for linkage using a nonparametric linkage (NPL) test is P=0.023.

If we tabulate allele counts at the marker, we see that the "5" allele is only seen in cases.

D17S74 Allele	1	2	3	4	**5**	6	7	8
Breast Cancer	1	2	0	1	**6**	1	0	1
Unaffected Female	0	1	3	4	**0**	1	0	3

The Pearson X^2 = 13.1, df=6, P=0.041

By contrast, the gene-dropping Monte-Carlo P-value is P=0.0002 (this family is segregating the c.2800 AA deletion).



Pearson Chi-square=13.1

Naive P-value=0.04

Empirical P-value=0.0002

Corrected χ^2 test of Bourgain *et al* (2003)

Bourgain *et al* (2003) suggest a modified form of the usual Pearson χ^2 test for allelic association, using the Fisher information assuming the between-individual correlations in allelic score and trait value all conform to an additive genetic model:

$$\frac{(p_{Case} - p)^2}{{}^1\!/_2\, p(1-p)K},$$

where,

$$K = \frac{1}{n_c^2} 1_c{}^T \Phi 1_c - \frac{2}{nn_c^2} 1^T \Phi 1_c + \frac{1}{n^2} 1^T \Phi 1$$

Φ is the numerator relationship matrix,

1_c is the trait indicator vector (1’s and 0’s) .

Numerator relationship matrix

The entries of **Φ** are the expectations that an allele present in the pair of specified relative is of the same ancestral origin (*identical by descent*).

Type of Relative Pair	Coefficient of Relationship
Self	1
Monozygotic Twins	1
Siblings/Dizygotic Twins	$^1/_2$
Parent-Offspring	$^1/_2$
Half-Siblings	$^1/_4$
Unilineal **kth** degree relative	$^1/_{2^k}$

Corrected χ^2 test of Bourgain *et al* (2003)

This corrected statistic is expected to be be distributed as χ^2_1.

In small pedigrees of the type seen in the BTNS, the P-values from this approach agree nicely with those coming from the gene-dropped simulation.

In larger pedigrees such as the terrier pedigree, there is less strong concordance between the methods. In the case of the highest ranked SNP arising from a genome-wide analysis of atopy in the terriers, the corrected χ^2 statistic value was 9.5 ($P=2.0\times10^{-3}$).

By constrast, the simulated P-value was 6×10^{-5} (95%CI=4.5×10^{-5}–7.9×10^{-5}).

In simulations with a similar missingness pattern, the variance of the corrected χ^2 is much higher than the simulation based test.

Corrected χ^2 test of Bourgain *et al* (2003)

Will this method be less successful in larger pedigrees?

- Estimation of p?
- Sensitivity of estimators to small genotyped N?

The population allele frequency p is (usually) estimated from the data at hand. If this is family data, estimation should allow for the relatedness: usually we estimate the frequency in the pedigree founders.

Bourgain *et al* (2003), McPeek *et al* (2004) suggest using an allele frequency estimator that is closely related to the corrected χ^2 method: a (quasi-likelihood) Best Linear Unbiased Estimator:

$$\hat{p} = (1^T \Phi^{-1} 1)^{-1} 1^T \Phi^{-1} G \,,$$

where $2G$ is the allele score vector for the marker.

This can occasionally be negative in large pedigrees if the true allele frequency is low.

MQLS test of Thornton and McPeek (2007)

Bourgain *et al* 2003 suggest some alternative tests, one of which they called the WQLS (weighted quasi-likelihood score). This weights the allele frequency estimate using the associated trait value correlation matrix (Φ), with for example

$p_c = \frac{w_c G}{w_c 1}$, with $w_c = \Phi^{-1} 1_c$, and a different variance correction,

$$K = (1_c^T \Phi^{-1} 1_c)(1_c^T \Phi^{-1} 1)^{-2} - (1^T \Phi^{-1} 1)^{-1}$$

They prefer this statistic over the corrected χ^2 for theoretical reasons, showing that it is locally most powerful under the test assumptions. In the simulations and examples they report (and in Thornton and McPeek 2007), however, the corrected χ^2 seems to actually fare as well.

MQLS test of Thornton and McPeek (2007)

The MQLS is an extension of the WQLS to use imputed genotypes in individuals with trait information but missing genotypes, as well as missing trait information and observed genotypes (an estimate of the trait prevalence is used for the latter group).

For the peak dog SNP, the WQLS=6.91,

and the MQLS:

Assumed prevalence	MQLS	P-value
0.2	9.18	0.002
0.4	13.55	2.3×10^{-4}
0.6*	16.14	5.9×10^{-5}

GLMM for familial association

- Preferred approach of geneticists
- Can be slow (lots of integration)
- Ascertainment?
- Little software suitable for large pedigrees

For the problematic terrier pedigree SNP, the MCMC algorithm in Sib-pair gives a Wald test P=1.1 $\times 10^{-3}$ using either observed or imputed genotypes.

APM for familial association

The affected pedigree member method was originally developed for linkage analysis (and has much in common with the random effects scoring approach of Commenges).

In its original formulation, the test statistic is based on identity-by-state similarity of cases (and controls, Ward (1993)), and so is sensitive to association as well as linkage. It is straightforward to gene-drop for the expectation and sampling variance of this statistic.

For the dog example, this obtains P-values of 2×10^{-4}.

Conclusion

In the case of the dog example, it is not clear to me whether I should regard this result as with following up, or completely consistent with a chance finding, given the large number of statistical tests carried out in a genome-wide analysis.