Practical exercises on transmission/disequilibrium(2): allelic association by parent-of-origin

David Clayton and Joanna Howson

6 Parent-of-origin

This next exercise examines the effect of parent-of-origin of alleles. If the MHC data are not already in memory, read them in again from the **Read Stata data set** menu, or with

```
. use mhc3iddm
```

(Don't forget to run the menus, `gamenu` if they aren't already running.) You can look at transmission of maternal and paternal alleles separately using the `tdt` command e.g. select **TDT** from the menus, choose alleles `bat2_1`, `bat2_2` and click the 'Mother' button on under 'Parental origin'. Click OK. Do the same again but click on the 'Father' button under parental origin. You could also have used the command lines,

```
. tdt bat2_*, mo
. tdt bat2_*, po
```

These results are suggestive of a difference. The `gtrr` command also has an option to track parent-of-origin of alleles. With this option in effect, the program distinguishes between genotypes `1|2` and `2|1`. Having chosen the **Genotype RR** menu, select `bat2_1`, `bat2_2` as the alleles, and click the 'Preserve parental origin of alleles' button on. Click OK.

```
. gtrr bat2_*, po
```

By default, `gtrr` calculates relative risks relative to the most frequent genotype. Here, however, that does not allow us easily to test for a parent-of-origin effect since we would need to assess the difference between the RR's for `1|2` and `2|1`. We can avoid the difficulty by choosing one of these as "reference" category for RR calculations e.g. under the **Genotype RR** menu, select `bat2_1` and `bat2_2` as alleles, click the parental origin button on, and this time fill in the 'Reference allele for relative risks' field with, `1|2`. Or use,

```
. gtrr bat2_*, po ref(1|2)
```

Is the effect of parent-of-origin statistically significant?

An alternative analysis is to create a case–control dataset and use conditional logistic regression. The `phase` option in the `pseudocc` command preserves the parent-of-origin of alleles, putting the maternal allele first and the paternal allele second. Choose **Pseudo-CC** from the menu and select `bat2_1`, `bat2_2` as alleles. Ensure the 'Preserve parental origin' button is clicked on, before clicking OK. The command line,

```
. pseudocc  bat2_*, po saving(ccmhc)
```

achieves the same thing. Read in the Stata data file, `ccmhc.dta`, you have just created.

The maternal and paternal alleles are now stored in `bat2_1` and `bat2_2` respectively. A conditional logistic regression analysis to fit separate effects for maternal and paternal alleles is

```
. clogit case bat2_1 bat2_2, group(set)
. test bat2_1 = bat2_2
```

Or choose the **Fit** menu and select `clogit` as the regression command, `case` as the response variable, `bat2_1` and `bat2_2` as the metric variables and put `group(set)` as the options. This is not (quite) the same answer as you obtained previously. The reason is that this is a slightly different model; here we have assumed that the RR for the homozygous (`2|2`) genotype is the product of effects of the maternal and paternal `2` alleles, while our previous analysis made no such assumption. We could relax this assumption in the regression analysis by allowing for an "interaction" between the alleles, but this is a more advanced use of regression.

Interaction between maternal genotype and child's genotype can masquerade as parent-of-origin effects. To test whether this is the case, we could stratify the analysis by maternal genotype. However, this would require us to link the maternal genotype data, contained on our original data file, with the current "case-control" data. Although this is straightforward to do (using the `merge` command), it is a little lengthy to do here and we shall leave it as an advanced exercise. It is, however, instructive since it demonstrates that it is extremely difficult to discriminate between mother×child interaction and parent-of-origin effects.

7 What is `pseudocc` doing?

It is important to understand precisely what `pseudocc` is doing here by returning to the example of Exercise 1. You should recreate the case-control dataset, this time preserving parent-of-origin of alleles. Read in the Stata file you created (`use exercise`). Select **Pseudo-CC** from the **TDT etc** sub menu. Click on `L1_1` and `L1_2` as the alleles; click the 'Preserve parental origin' button on and save the data as `casecon`. Click OK and read in the data set, casecon.dta. Alternatively use the command lines,

```
. pseudocc L1_*, po saving(casecon) replace
. use casecon
```

Again you should use the data editor to browse the file that has been created. Pay particular attention to the action that has been taken for families (4) and (5). Why do you think different action has been taken in these two cases?

8 Counting case-parent triads

Before proceeding we must first restore the original data to memory:

```
. use mhc3iddm
```

An alternative approach to the problem of parent–of–origin effects has been developed by Weinberg and colleagues (*American Journal of Human Genetics*, **65**:229-235, 1998). This involves counting triads of parents and affected offspring. A tool to do this in Stata is `trios`. For the `bat2` marker:

```
. trios bat2*, first
```

Note the use of the `first` option so that only the first affected offspring of any family is used. This is used because Weinberg's method may be misleading if there are multiple affected offspring.[1]

There are 15 types of triad.

1. Identify the 10 types of triad in which maternal and paternal genotypes are different
2. Arrange these into five pairs such that the two members of each pair are the same except for reversal of paternal and maternal genotype. You can check that you have done this correctly, by using the command

   ```
   trios bat2*, first po
   ```

3. Were it not for selection of triads by affected offspring we would expect the two frequencies within a pair to be equal. However, because of the selection, we expect the ratio of frequencies to reflect the ratio of offspring risks.
4. For each pair, how would you expect the ratio of frequencies to be affected by different risks being associated with paternal and maternal copies of the '2' allele?
5. Informally, do the triad counts suggest such an effect?

More formally, we can carry out a test for this using the command `origin`. This command fits the logistic regression model which predicts the ratio of frequencies of the two triads in each pair

```
. origin bat2*
```

In this case the logistic regression model is a very simple one — it is required only to detect deviation from a 50:50 split in all pairs of triads. However, Weinberg also pointed out that a direct (presumably inter-uterine) effect of maternal genotype on subsequent disease risk in the offspring may also distort the ratio of triad frequencies. This may be allowed for in the multiple regression;the analysis can conveniently be carried out as follows:

```
. origin bat2*, mat
```

[1] Even with this option there may be difficulties. If families have been ascertained to have at least two affected offspring and there is a parent-of-origin effect, the expected distribution of triads will not be the same as predicted by the simple theory.

9 Extending the case/pseudo–control analysis

Weinberg's approach is more efficient than the case/pseudo–control approach descibed earlier. Essentially, the additional information is derived at the expense of an additional assumption — that of *exchangeability* of parental genotypes. Formally this means that for a given *mating type*, $a/b + c/d$ say, in the population it is just as likely that the father is a/b and the mother is c/d as if the mother is a/b and the father c/d. This assumption may also be incorporated into the creation of a case/pseudo–control study; additional pseudo–controls are created by simply switching the maternal and paternal alleles of the existing case and pseudo–controls.

To experiment with this, return to the `exercise.dta` datafile. You can create the new case/pseudocontrol study from the menu interface or, from the command line, as follows:

```
. pseudocc L1_*, first phase exch saving(casecon) replace
. use casecon
```

You should repeat the earlier analysis with this new, extended case/pseudo–control data.