ALBERT Documentation

Version 1.0
Last Updated: 2006/05/01

Table of Contents

1 Introduction

1.1 What is ALBERT?

ALBERT, A Likelihood-Based Estimation of Risk in Trios, is a `C` program that estimates genotype relative risks, genotyping error rates and population risk allele frequencies from marker genotype data in case-parent trios. ALBERT uses the distribution of trio marker genotypes to compute maximum likelihood estimates for the parameters.

IMPORTANT: Do not remove Mendelian inconsistencies from your data before running ALBERT! The program makes use of visible genotyping errors in its estimates of error rate, genotype relative risks and allele frequency.

Until ALBERT is published, only the executable file will be available. Following publication, the source code will also be available. If you use ALBERT, please cite

```
Mitchell AA and Thompson EA.
  ALBERT: A Likelihood-Based Estimation of Risk in Trios.  Submitted
```

Email questions or bugs to Adele Mitchell at `adele@u.washington.edu`.

1.2 Coming soon

Several improvements to ALBERT are coming soon.

1. In the not-too-distant future, a Windows executable will be available for ALBERT.
2. At present, if any member of a trio is missing genotype data at a marker, that trio will not be included in the analysis at that marker. Work is in progress to model missing data so that all available data can be used.
3. ALBERT assumes that the trios in a dataset were drawn from a single population in Hardy-Weinberg equilibrium at each marker locus. Violation of this assumption can lead to an increased false positive rate. A future version of ALBERT will model population subdivision to avoid this problem.

2 Input Files

Two files are required when running ALBERT; these are a marker file and a pedigree file.

2.1 Marker File

The marker file must contain a single column with one marker name per row. The maximum length of each marker name is 20 alphanumeric characters. If the number of markers listed in the marker file is inconsistent with the number of columns in the pedigree file, ALBERT will terminate and an error message will be sent to the 'ALBERT.out' file.

2.2 Pedigree File

The pedigree file contains family relationship information and marker genotype data. The file must be in Standard LINKAGE file format with entries separated by tabs or spaces. Use one line per individual with columns for family ID, individual ID, father's ID, mother's ID, sex and trait followed by genotype data. The total number of columns will be six plus two times the number of markers. A sample pedigree file is below.

```
1  1  0  0  1  1  4  4  2  3  1  3
1  2  0  0  2  1  3  4  2  2  1  1
1  3  1  2  1  2  3  4  2  2  1  3
2  1  0  0  1  1  4  4  2  3  1  1
2  2  0  0  2  1  3  3  2  3  1  3
2  3  1  2  1  2  3  4  2  3  1  3
```

Guidelines for pedigree file:

- All entries in the pedigree file must be integers
- Make sure that each family ID number is unique
- Individual ID numbers must be unique within the family
- Code all missing values as '0' (zero)
- Currently, sex is not used in the program, but the column must be present. Sex can be coded with any integer values.
- Code the trait value as '0' for missing, '1' for unaffected and '2' for affected
- ALBERT will use at most one affected child per family. If more than one child in a family is coded as 'affected', the one with the lowest ID number will be used in the analysis. Do not break up families with more than one affected individual into smaller families, as the method assumes affected individuals are unrelated to one another.

3 Running ALBERT

ALBERT is run from the unix command line with *albert* followed by three required arguments. The format is:

```
albert marker_file pedigree_file output_file error_model
```

If either of the marker and pedigree files does not exist, ALBERT will terminate. The output of the program will be written to the output file. A new file will be created if it does not already exist. If the file does exist, results from the current run will be appended to the exisiting file. The error model argument must be either '`1`' or '`2`', for the simple error model and the allelic dropout error model, respectively. Error models are described in the Methods section. To run ALBERT with a marker file called '`markers.txt`', a pedigree file called '`data.txt`', sending results to the output file '`out.txt`', using the allelic dropout error model:

```
albert markers.txt data.txt out.txt 2
```

4 Output Files

Two output files are created each time ALBERT runs. The first file, '`ALBERT.out`', is generated automatically. It contains error messages and notations about missing data. The second file is named by the user on the command line when running the program. This file contains the results of the run. A sample is below.

```
Marker  TDTp   allele  p     e      r1   r2    nullP  nullE   LRT   pvalue
rs2742   2.3e-01  3    0.57  0.035  1.21  1.53  0.60   0.0363  0.88  6.45e-01
rs3026   1.1e-01  1    0.75  0.028  0.99  1.59  0.80   0.0338  1.89  3.90e-01
rs3027   1.8e-01  4    0.81  0.027  0.80  1.21  0.84   0.0305  1.39  5.00e-01
rs7417   1.5e-05  3    0.58  0.006  2.63  7.47  0.70   0.0105  18.91 7.84e-05
rs2505   1.0e-04  2    0.69  0.011  1.68  4.69  0.78   0.0211  12.53 1.90e-03
```

The columns in the output file are interpreted as follows:

- TDTp: p-value obtained using the transmission disequilibrium test (TDT) (Spielman et al., Am J Hum Genet 1993 52:506-16)
- allele: This is the allele associated with higher risk of disease
- p: Maximum likelihood estimate (MLE) of population frequency of allele in previous column for the model in which genotype relative risks are allowed to vary (i.e., the alternative hypothesis).
- e: MLE of genotyping error rate under alternative hypothesis. In this example, the simple error model was chosen so there is only one error parameter. When the allelic dropout error model is chosen, there will be two parameters.
- r_1: MLE of genotype relative risk associated with carrying 1 copy the allele in column three vs. no copies.
- r_2: MLE of genotype relative risk associated with carrying 2 copies the allele in column three vs. no copies.
- nullP: MLE of population frequency of allele in column three for the model in which genotype relative risks are fixed at 1.0 (i.e., the null hypothesis).
- nullE: MLE of genotyping error rate under the null hypothesis.
- LRT: 2 times (LL minus nullLL)
- pvalue: ALBERT pvalue. LRT using chi square distribution with 2 degrees of freedom. The p value not adjusted for multiple testing.

If ALBERT estimates non-zero error rates at all or nearly all markers, the dataset may contain individuals whose relationships are misspecified, such as with non-paternity or sample mix-up. ALBERT's false positive rate will not be inflated if this is true, but power may be lost and parameter estimates may be off. Therefore, we advise running a pedigree checking program to identify and remove such individuals before running ALBERT. Do not however, blank out genotypes that are Mendelian-inconsistent if you do not believe the relationship to be misspecified, as the program uses these inconsistencies in parameter estimation.

5 Methods

ALBERT tests for association between a biallelic marker and a dichotomous trait. Parameters estimated are p, r_1, r_2 and e (or e_1 and e_2), as described in the previous section. Two different error models are available in the current implementation of ALBERT; these are the simple error model and the allelic dropout error model. In the simple model, a single error parameter, e, is used. e is an allele-wise error rate, representing the probability that one allele is miscalled as the other. The allelic dropout model uses two parameters, e_1 and e_2, which represent the probability of miscalling a homozygote as a heterozygote and vice versa, respectively. ALBERT also performs a likelihood ratio test, comparing the model in which r_1 and r_2 are free to one in which $r_1 = r_2 = 1.0$.

To do this, we calculate the probability of observing any possible parent-child genotype combination using one of the two error models and assuming that the trio was ascertained through a single affected offspring SAO. To model SAO ascertainment, we assume HWE at the marker locus. Then, the probability that an affected individual has marker genotype AA, Aa or aa is $p^2 r_2/R$, $2p(1-p)r_1/R$ or $(1-p)^2/R$, where p is the population frequency of the '`A`' allele, r_1 and r_2 are the genotype relative risks for the '`AA`' and '`Aa`' genotypes, relative to the '`aa`' genotype, and $R = p^2 r_2 + 2p(1-p)r_1 + (1-p)^2$.

Once we have the distribution of affected children's genotypes, we find the distribution of parental genotypes conditional on the child's. To do this, we assume random mating and HWE. We can then find the joint distribution of mother, father and child's genotypes (MFC). $Pr(MFC) = \Sigma_{TC} Pr(TP \mid TC) Pr(TC)$, where TP is the true parental genotypes and TC is the true child genotype.

To allow for genotyping error, we compute the probability of observing any particular parental genotypes given their true genotypes and any particular child genotype given the child's true genotype. These are represented as $Pr(OP \mid TP)$ and $Pr(OC \mid TC)$, respectively, where OP and OC are the observed parent and child genotypes in a trio.

Finally, combining the above factors, we have
$Pr(OP,\ OC \mid SAO) = \Sigma_{TP}\ \Sigma_{TC}\ Pr(OP \mid TP)\ Pr(OC \mid TC)\ Pr(TP,\ TC \mid SAO)$.

We then perform a likelihood ratio test using the maximum likelihood estimates of the parameters under the alternative hypothesis, $\hat{\ }e, \hat{\ }r_1, \hat{\ }r_2, \hat{\ }p$, and under the null, $\tilde{e}, \tilde{p}$. The ALBERT output reports the MLEs under the null and alternative, the log likelihoods and the p-value for the likelihood ratio test.