GWAS, Quality Control and Family based analysis workshop February 2016 – Matthew McQueen
TUTORIAL 5 - Family-Based Association Analysis

Basic FBAT Analysis

The goal of this tutorial is to provide an overview of the basic commands in FBAT, as well as some of the available analysis options. This is by no means an exhaustive list of commands (see the FBAT toolkit for more comprehensive analysis options), rather, an introduction to the FBAT software. We will begin with looking at the AD1 data set. Recall the description of these data:

NOTE: These files are located in the *~/cbio2016/fbat* directory

Alzheimer's Disease (AD1) Dataset

This data set is a sub sample from the NIMH Genetics Initiative AD sample. The ascertainment and assessment of AD families collected have been discussed in Blacker *et al.* (1997). None of the families in this data set have parental genotype information; practically all of them have both affected and unaffected offspring. In total there are 901 individuals contained in 301 nuclear families (about 3 sibs per family, with at least one affected.) The prevalence of disease is about 0.1 in this population.

Genotypes

The pedigree file, **AD1.ped**, contains genotype information for two candidate genes, apolipoprotein-E (APOE) and alpha-2-macroglobulin (A2M). The APOE gene is multi-allelic (alleles 2, 3 and 4), while the A2M gene is bi-allelic. The dataset also contains the affection status (2=affected, 1=unaffected, 0=missing). This is the only available phenotype.

Before we begin, let's create a new directory for our fbat output:

```
student@courses:~$ cd
student@courses:~$ mkdir fbatout
student@courses:~$ cd fbatout
```

Opening FBAT

Let's start by running FBAT interactively by typing `fbat` at the command prompt.

GWAS, Quality Control and Family based analysis workshop February 2016 – Matthew McQueen
TUTORIAL 5 - Family-Based Association Analysis

Dichotomous Trait Analysis

1. Creating a logfile

```
log AD1.log
```

A file, named "AD1.log" has now been created in the *~cbio2016/fbat* directory.

2. Loading the pedigree file

```
load /student_data/cbio2016/fbat/AD1.ped
```

```
>> load AD1.ped
read in: 2 markers from 308 pedigrees (301 nuclear families,901 persons)
```

3. Let's start by looking at both markers using the default settings.

```
fbat
```

```
>> fbat
trait affection; offset 0.000; model additive; test bi-allelic; minsize 10;
min_freq 0.000; p 1.000; maxcmh 1000

Marker  Allele   afreq   fam#     S-E(S)      Var(S)       Z           P
-------------------------------------------------------------------------
a2m     1        0.843     49    -14.367      16.335  -3.555    0.000378
a2m     2        0.157     49     14.367      16.335   3.555    0.000378
apoe    2        0.036     15     -7.209       4.269  -3.489    0.000484
apoe    3        0.533     85    -22.301      36.836  -3.674    0.000238
apoe    4        0.430     83     29.511      37.131   4.843    1.28e-06
```

Let's review the output…

Note that the defaults are in effect here: trait is affection as defined in the ped file. Model is additive, and offset is zero. We'll see how to change those defaults later. **afreq** is the frequency of the allele, **#fam** is the number of informative families. For more information on **S, E(S),** and **Var(S)**, see the FBAT toolkit and refer to lecture notes. In the lecture notes, U = S-E(S). **Z** is the value of the FBAT statistic (which asymptotically follows a standard normal distribution) and **P** is the corresponding **p-value** (2-sided) from the test statistic. As expected, apoe allele 4 is highly significant, associated with an increased risk of AD.

Notice that for the additive model, and a bi-allelic marker (a2m), the tests for the two alleles are identical, except with different sign. With more than two alleles (apoe), the tests will be different for each allele.

4. Exploring other genetic models

Type **`model r`** to see what the recessive model shows:

```
>> model r
current genetic model is recessive
```

To look at a2m only, type **`fbat a2m`**. Note that fbat is case sensitive for trait and marker names.

```
>> fbat a2m
trait affection; offset 0.000; model recessive; test bi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000

Marker Allele   afreq   fam#     S-E(S)      Var(S)       Z           P
-------------------------------------------------------------------------
a2m    1        0.843     49    -14.367      16.335  -3.555    0.000378
a2m    2        0.157     49     14.367      16.335   3.555    0.000378
```

Note that the two tests are now different. By typing model d, you will see that the recessive model for allele 1 gives the same answer as the dominant model for allele 2, and vice-versa.

5. Multiallelic Tests (Multiple Degrees of Freedom)

To get the global genotype test (for any effect of any genotype) type **`mode m`**

```
>> mode m
current test mode is multi-allelic

>> fbat apoe
trait affection; offset 0.000; model additive; test multi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000

Marker Allele#   Fam#     DF       CHISQ           P
-------------------------------------------------
apoe         3     90      2      30.507    2.37e-07
```

This is a 2 DF test which compares all genotype frequencies to their expectations.

6. Optimizing the offset

Since the prevalence of the disorder is known to be about 0.1 in this population, we will use this for an offset. Type **`offset 0.1`** and then **`fbat apoe`**.

```
>> offset 0.1
current trait offset is 0.100000
>> fbat apoe
trait affection; offset 0.100; model additive; test multi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000

Marker Allele#   Fam#     DF       CHISQ           P
-------------------------------------------------
apoe         3     92      2      31.078    1.78e-07
```

The effect of using an offset is very marginal here. Because there are no parents in the dataset, many unaffected siblings are included in the dataset. We also try using `-o`:

Type **`fbat -o apoe`**

```
>> fbat -o apoe
trait affection; offset 0.100; model additive; test multi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000

Marker Allele#   Fam#     DF       CHISQ           P      Offset
-------------------------------------------------------------
apoe         3     92      2      28.332    7.04e-07       0.577
```

The offset is very large, because it is approximately the prevalence in the *sample.* With large samples and highly statistically significant results, we often will not see big differences in different offsets. However, subsequently we compare using an offset of 0.5 (or –o) with an offset of 0 for SNP 23 in the xbat dataset—here we see a big difference.

7. Empirical variance option

The chromosome 19 region that harbors the APOE gene has demonstrated linkage to AD. Therefore, we need to adjust our analyses to account for the linkage, and test for association in presence of linkage (H_{02}). However, in FBAT, we cannot use the '-o' and '-e' options simultaneously. One way around this is to assign the FBAT offset (determined in [5]) prior to analysis.

Type **`offset 0.01`** (assign an appropriate offset) and then type **`fbat -e apoe`**

```
>> offset 0.10
current trait offset is 0.100000
>> fbat -e apoe
trait affection; offset 0.100; model additive; test multi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000
```

```
Marker Allele#    Fam#      DF       CHISQ           P
-------------------------------------------------------
apoe          3      82       2      23.961    6.27e-06
```

Accounting for linkage results in a slightly larger p-value.

8. Close up the analysis
Type **`log off`** to close the log, and then **`quit`** to exit the program.

9. Affected Status.
First, restart fbat and then load up the new dataset

```
load /student_data/cbio2016/fbat/xbat.ped
```

We now briefly illustrate the offset command on SNP23 using affection status. To create affected status, a quantitative variable was dichotomized at the median, so the sample prevalence is 50%.
Type **`fbat SNP23`**. A reminder that fbat is case sensitive for trait and marker names.

```
>> fbat SNP23
trait affection; offset 0.000; model additive; test bi-allelic; minsize 10; min_freq
0.000; p 1.000; maxcmh 1000

Marker  Allele   afreq   fam#     S-E(S)      Var(S)       Z           P
-------------------------------------------------------------------------
SNP23   2        0.713    326    -14.000     105.500  -1.363    0.172877
SNP23   4        0.287    326     14.000     105.500   1.363    0.172877
```

With an offset of zero, only affected are used.

Type **`fbat -o SNP23`**

```
>> fbat -o SNP23
trait affection; offset 0.000; model additive; test bi-allelic; minsize 10; min_freq 0.000; p
1.000; maxcmh 1000

Marker  Allele   afreq   fam#     S-E(S)      Var(S)       Z           P      Offset
-------------------------------------------------------------------------------------
SNP23   2        0.713    664    -18.988      52.875  -2.611    0.009019       0.499
SNP23   4        0.287    664     18.988      52.875   2.611    0.009019       0.499
```

Including the unaffected has substantially increased the power.

10. Close up the affected trait analysis
Type **`log off`** (but do not quit).

GWAS, Quality Control and Family based analysis workshop February 2016 – Matthew McQueen
TUTORIAL 5 - Family-Based Association Analysis

Quantitative Trait Analysis

1. Start a new log file
```
log qtl.log
```

2. Load the phenotype file
```
load /student_data/cbio2016/fbat/xbat.phe
```

3. Select the trait for analysis
```
trait QTL1
```

4. Select the genetic model for analysis (additive model)
```
model a
```

5. Specify the offset
Again, we want to recode the trait with an offset for optimal power. For quantitative traits, the value that maximizes the power is the population mean for the trait (or sample mean if ascertainment is random). For this example, we will let FBAT choose the optimal offset as this is an unascertained sample.

6. Test SNPs 1, 5 and 10 for association
```
fbat –o SNP1 SNP5 SNP10
```

```
>> fbat -o SNP1 SNP5 SNP10
trait QTL1; offset 0.000; model additive; test bi-allelic; minsize 10; min_freq 0.000; p 1.000;
maxcmh 1000

Marker  Allele   afreq   fam#     S-E(S)      Var(S)       Z           P      Offset
-------------------------------------------------------------------------------------
SNP1    1        0.449    747    184.299   62179.518   0.739    0.459852     120.200
SNP1    3        0.551    747   -184.299   62179.518  -0.739    0.459852     120.200
SNP5    1        0.583    728   -208.906   62611.051  -0.835    0.403784     120.497
SNP5    3        0.417    728    208.906   62611.051   0.835    0.403784     120.497
SNP10   2        0.651    697   -109.045   63119.618  -0.434    0.664265     121.763
SNP10   4        0.349    697    109.045   63119.618   0.434    0.664265     121.763
```

Note that the offset is close to the population mean, but varies slightly for each SNP.

7. Close up the quantitative trait analysis
```
log off
quit
```