TDT vignette
Use of snpMatrix in family–based studies

David Clayton

February 4, 2011

Pedigree data

The `snpMatrix` package contains some tools for analysis of family-based studies. These assume that a subject support file provides the information necessary to reconstruct pedigrees in the well-known format used in the *LINKAGE* package. Each line of the support file must contain an identifier of the *pedigree* to which the individual belongs, together with an identifier of subject within pedigree, and the within-pedigree identifiers for the subject's father and mother. Usually this information, together with phenotype data, will be contained in a dataframe with rownames which link to the rownames of the `snp.matrix` containing the genotype data. The following commands read some illustrative data on 3,017 subjects and 43 (autosomal) SNPs[1]. The data consist of a dataframe containing the subject and pedigree information (`pedfile`) and a `snp.matrix` containing the genotype data (`genotypes`):

```
> require(snpMatrix)
> data(families)
> head(genotypes)
```

```
A snp.matrix with  6 rows and  43 columns
Row names:  id02336 ... id01069
Col names:  rs91126 ... rs98918
```

```
> head(pedfile)
```

```
        familyid member father mother sex affected
id02336  fam0005      1     NA     NA   1        1
id00695  fam0005      2     NA     NA   2        1
id02750  fam0005      3      1      2   2        2
id01836  fam0005      4      1      2   2        2
id02533  fam0006      1     NA     NA   2        1
id01069  fam0006      2     NA     NA   1        1
```

[1]These data are on a much smaller scale than would arise in genome-wide studies, but serve to illustrate the available tools. Note, however, that execution speeds are quite adequate for genome-wide data.

The first family comprises four individuals: two parents and two sibling offspring. The parents are "founders" in the pedigree, *i.e.* there is no data for their parents, so that their `father` and `mother` identifiers are set to `NA`. This differs from the convention in the *LINKAGE* package, which would code these as zero. Otherwise coding is as in *LINKAGE*: `sex` is coded 1 for male and 2 for female, and disease status (`affected`) is coded 1 for unaffected and 2 for affected.

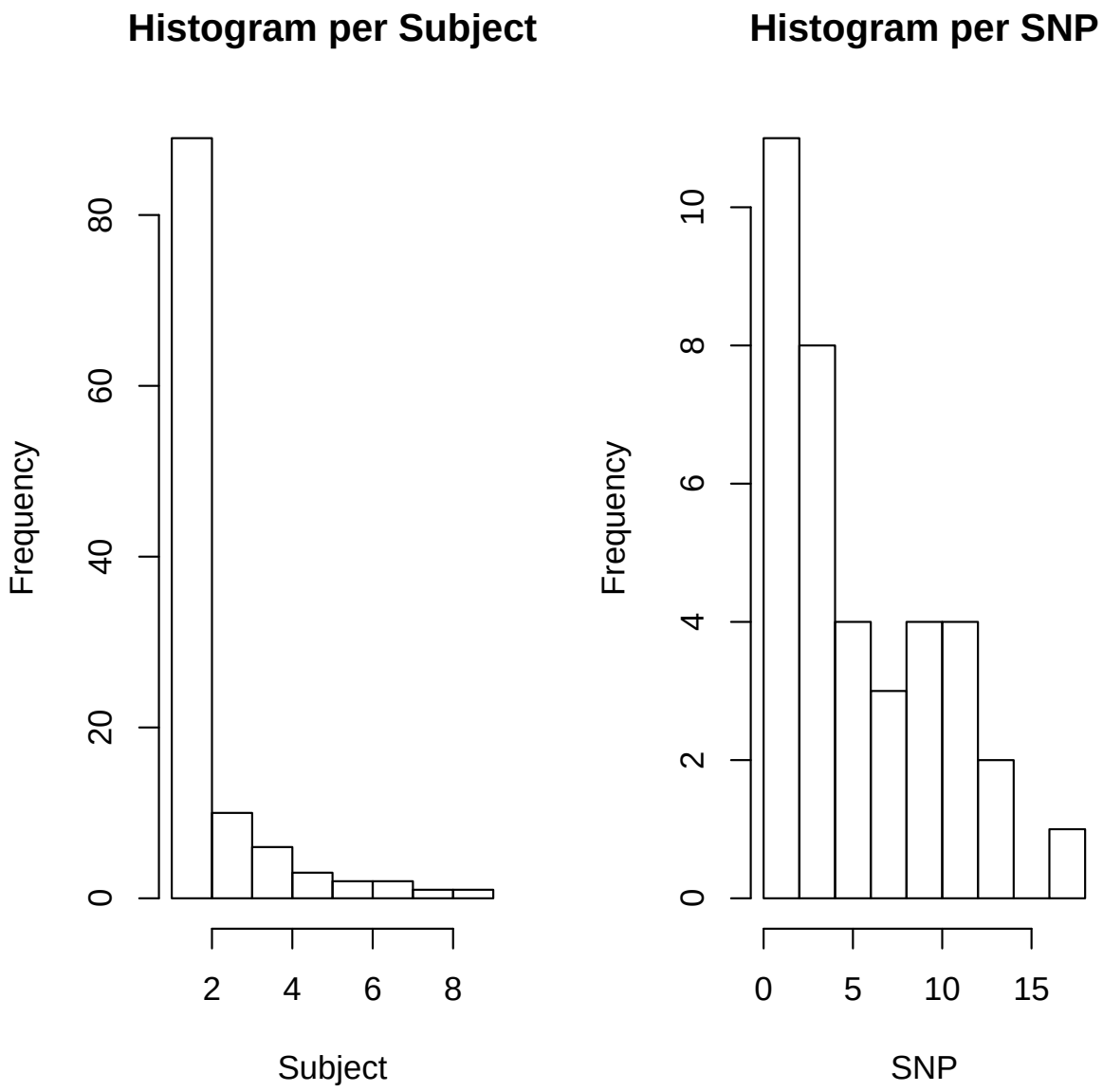
Checking for mis-inheritances

The function `misinherits` counts non-Mendelian inheritances in the data. It returns a logical matrix with one row for each subject who has any mis-inheritances and one column for each SNP which was ever mis-inherited.

```
> mis <- misinherits(data = pedfile, snp.data = genotypes)
> dim(mis)

[1] 114  37
```

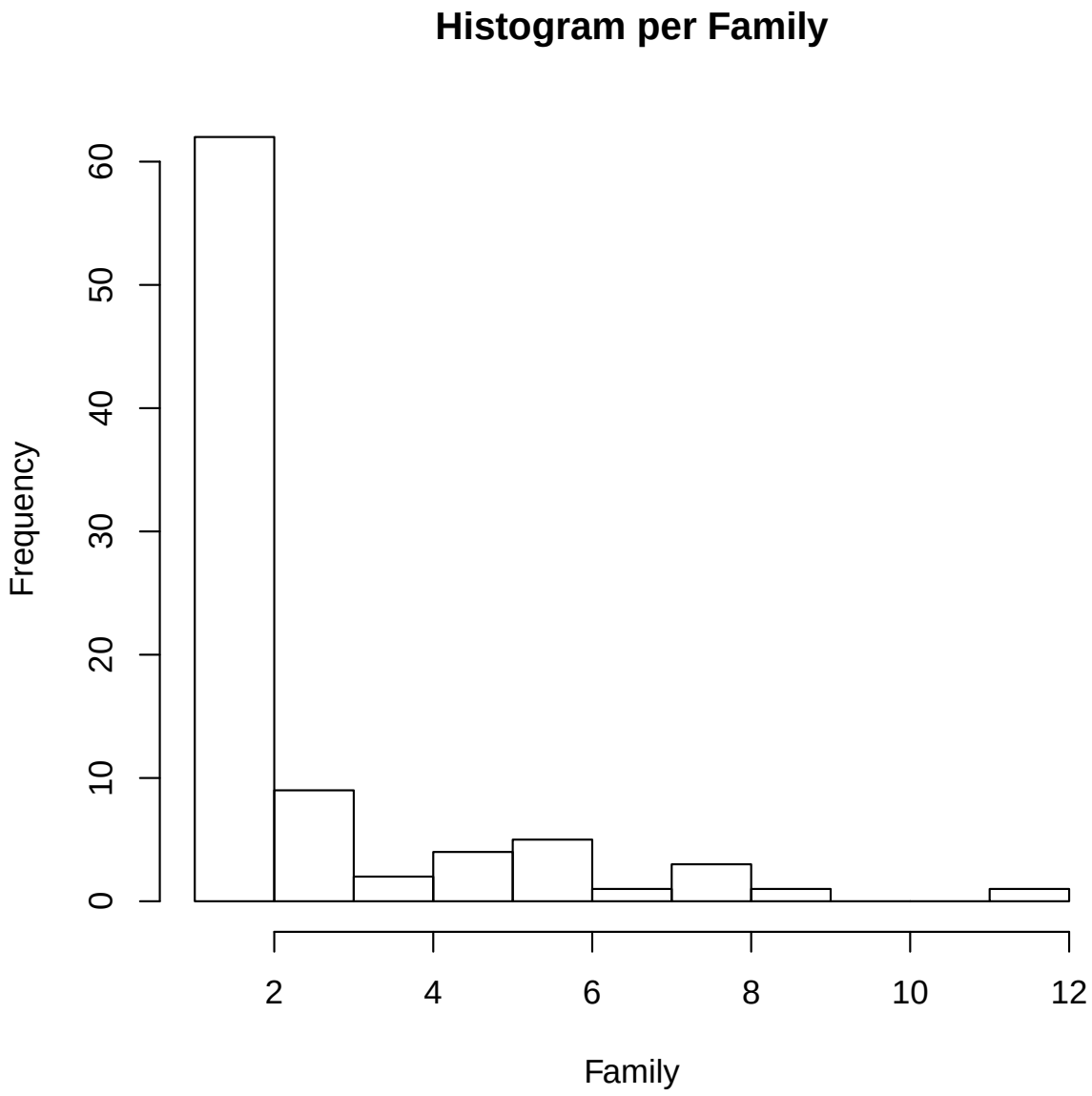
Thus, 114 of the subjects and 37 of the SNPs had at least one mis-inheritance. The following commands count mis-inheritances per subject and plot its frequency distribution, and similarly, for mis-inheritances per SNP:

```
> per.subj <- apply(mis, 1, sum, na.rm = TRUE)
> per.snp <- apply(mis, 2, sum, na.rm = TRUE)
> par(mfrow = c(1, 2))
> hist(per.subj, main = "Histogram per Subject", xlab = "Subject")
> hist(per.snp, main = "Histogram per SNP", xlab = "SNP")
```



Note that mis-inheritances must be ascribed to offspring, although the error may lie with the parent data. The following commands first extract the pedigree identifiers for mis-inheriting subjects and go on to chart the numbers of mis-inheritances per family:

```
> fam <- pedfile[rownames(mis), "familyid"]
> per.fam <- tapply(per.subj, fam, sum)
> par(mfrow = c(1, 1))
> hist(per.fam, main = "Histogram per Family", xlab = "Family")
```



None of the above analyses suggest serious problems with the data, although there are clearly a few genotyping errors.

TDT tests

At present, the package only allows testing of discrete disease phenotypes in case–parent trios — basically the Transmission/Disequilibrium Test (TDT). This is carried out by the function `tdt.snp`, which returns the same class of object as that returned by `single.snp.tests`; allelic (1 df) and genotypic (2 df) tests are computed. The following commands compute

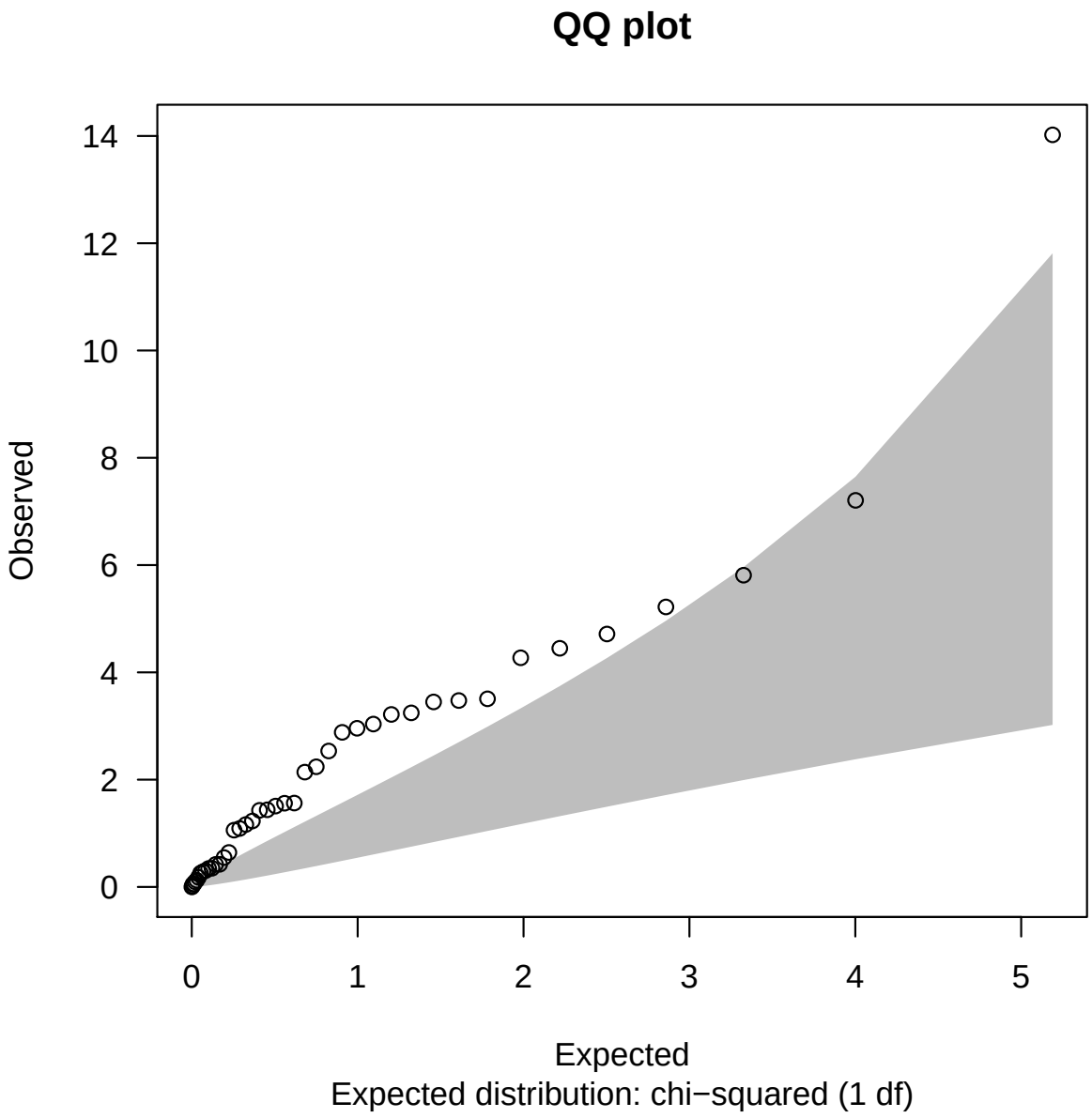
the tests, display the p-values, and plot quantile–quantile plots of the 1 df tests chi-squared statistics:

```
> tests <- tdt.snp(data = pedfile, snp.data = genotypes)

Analysing 1466 potentially complete trios in 733 different pedigrees

> cbind(p.values.1df = p.value(tests, 1),
+       p.values.2df = p.value(tests, 2))

        p.values.1df p.values.2df
rs91126 0.3034836640 0.3840474189
rs62927 0.1113713350 0.1069552818
rs79960 0.6720969384 0.3559056330
rs19348 0.0895550744 0.1933553084
rs99786 0.0072618187 0.0244219016
rs36984 0.1434326196 0.1951188651
rs52628 0.8906175462 0.3034008949
rs6699  0.0001807362 0.0006286452
rs12373 0.4590596257 0.6928644074
rs35215 0.2115224294 0.1712371429
rs41229 0.0159202669 0.0507959278
rs86267 0.1344540153 0.0815662213
rs23261 0.5942123774 0.2669689090
rs69208 0.0854324416 0.2252305755
rs16483 0.6120898801 0.6957040098
rs8558  0.5159360077 0.2326343059
rs55762 0.0632861527 0.1362184523
rs8124  0.2111053457 0.4560817439
rs72056 0.0298913543 0.0936747901
rs82369 0.0813983946 0.1726193036
rs97686 0.5809872358 0.5346901234
rs77065 0.7236736098           NA
rs53106 1.0000000000 0.0443213934
rs37378 0.2194915577 0.2861004147
rs83832 0.8142257039 0.8868568605
rs35431 0.4226780742 0.2989153259
rs61158 0.5167185935 0.7750913645
rs32410 0.0387409847 0.1104896192
rs85906 0.2319977236 0.2760623003
rs83977 0.2807488029 0.3020104520
rs24527 0.2963306800 0.5462696091
rs73721 0.0729239892 0.0373514963
rs36088 0.0349165828 0.1081036861
```

```
rs32998 0.5571397270 0.8361802683
rs5566  0.0716136366 0.1474327754
rs98256 0.5549898129 0.5124651432
rs29479 0.8193227772 0.8139000784
rs42938 0.0611009304 0.1616517671
rs32018 0.7572705888 0.7923054438
rs39483 0.2304232228 0.2949232314
rs42367 0.2674484200 0.5321903430
rs87640 0.0223275596 0.0690116962
rs98918 0.0622805951 0.1050760176
```

```
> qq.chisq(chi.squared(tests, 1), df = 1)
```

```
        N   omitted    lambda
43.000000  0.000000  3.454497
```



Since these SNPs were all in a region of known association, the overdispersion of test statistics is not surprising. Note that, because each family had two affected offspring, there were twice as many parent-offspring trios as families. In the above tests, the contribution of the two trios in each family to the test statistic have been assumed to be independent. When there is *linkage* between the genetic locus and disease trait, this assumption is incorrect and an alternative variance estimate can be used by specifying `robust=TRUE` in the call. However, in practice, linkage is very rarely strong enough to require this correction.