Imputed SNP analyses and meta-analysis with snpStats

David Clayton

April 29, 2025

Getting started

The need for imputation in SNP analysis studies occurs when we have a smaller set of samples in which a large number of SNPs have been typed, and a larger set of samples typed in only a subset of the SNPs. We use the smaller, complete dataset (which will be termed the *training dataset*) to impute the missing SNPs in the larger, incomplete dataset (the *target dataset*). Examples of such applications include:

- use of HapMap data to impute association tests for a large number of SNPs, given data from genome-wide studies using, for example, a 500K SNP array, and
- meta-analyses which seek to combine results from two platforms such as the Affymetrix 500K and Illumina 550K platforms.

Here we will not use a real example such as the above to explore the use of `snpStats` for imputation, but generate a fictitious example using the data analysed in earlier exercises. This is particularly artificial in that we have seen that these data suffer from extreme heterogeneity of population structure.

We start by attaching the required libraries and accessing the data used in the exercises:

```
> library(snpStats)
> library(hexbin)
> data(for.exercise)
```

We shall sample 200 subjects in our fictitious study as the training data set, select every second SNP to be missing in the target dataset, and split the training set into two parts accordingly:

```
> training <- sample(1000, 200)
> select <- seq(1, ncol(snps.10),2)
> missing <- snps.10[training, select]
> present <- snps.10[training, -select]
> missing
```

```
A SnpMatrix with  200 rows and  14251 columns
Row names:  ceu.480 ... jpt.92
Col names:  rs7909677 ... rs12218790
```

```
> present
```

```
A SnpMatrix with  200 rows and  14250 columns
Row names:  ceu.480 ... jpt.92
Col names:  rs7093061 ... rs7899159
```

Thus the training dataset consists of the objects `missing` and `present`. The target dataset holds a subset of the SNPs for the remaining 800 subjects.

```
> target <- snps.10[-training, -select]
> target
```

```
A SnpMatrix with  800 rows and  14250 columns
Row names:  jpt.869 ... jpt.347
Col names:  rs7093061 ... rs7899159
```

But, in order to see how successful we have been with imputation, we will also save the SNPs we have removed from the target dataset

```
> lost <- snps.10[-training, select]
> lost
```

```
A SnpMatrix with  800 rows and  14251 columns
Row names:  jpt.869 ... jpt.347
Col names:  rs7909677 ... rs12218790
```

We also need to know where the SNPs are on the chromosome in order to avoid having to search the entire chromosome for suitable predictors of a missing SNP:

```
> pos.miss <- snp.support$position[select]
> pos.pres <- snp.support$position[-select]
```

Calculating the imputation rules

The next step is to calculate a set of rules which for imputing the `missing` SNPs from the `present` SNPs. This is carried out by the function `snp.imputation`[1]:

```
> rules <- snp.imputation(present, missing, pos.pres, pos.miss)
```

[1]Sometimes this command generates a warning message concerning the maximum number of EM iterations. If this only concerns a small proportion of the SNPs to be imputed it can be ignored.

```
SNPs tagged by a single SNP: 4901
SNPs tagged by multiple tag haplotypes (saturated model): 9180
```

This step executes remarkably quickly when we consider what the function has done. For each of the 14251 SNPs in the "missing" set, the function has performed a forward step-wise regression on the 50 nearest SNPs in the "present" set, stopping each search either when the R^2 for prediction exceeds 0.95, or after including 4 SNPs in the regression, or until R^2 is not improved by at least 0.05. The figure 50 is the default value of the `try` argument of the function, while the values 0.95, 4 and 0.05 together make up the default value of the `stopping` argument. After the predictor, or "tag" SNPs have been chosen, the haplotypes of the target SNP plus tags was phased and haplotype frequencies calculated using the EM algorithm. These frequencies were then stored in the `rules` object. [2]

A short listing of the first 10 rules follows:

```
> rules[1:10]

rs7909677  ~  rs2496276+rs4881551+rs4880750+rs9419498 (MAF = 0.0530303, R-squared = 0.86
rs12773042  ~  rs2496276+rs4881551+rs4880750+rs9419498 (MAF = 0.05076142, R-squared = 0.
rs11253563  ~  rs7093061+rs9419496+rs7898275+rs2246654 (MAF = 0.251269, R-squared = 0.97
rs4881552  ~  rs7475011+rs4881551+rs7093061+rs2379078 (MAF = 0.3959391, R-squared = 0.97
rs10904596  ~  rs7093061+rs9419496+rs7898275+rs2246654 (MAF = 0.2424242, R-squared = 0.9
rs4880781  ~  rs4880983+rs4880750+rs10736957+rs3740304 (MAF = 0.201005, R-squared = 0.98
rs7910845  ~  rs3123252 (MAF = 0.04615385, R-squared = 1)
rs6560730  ~  rs9419496 (MAF = 0.3080808, R-squared = 0.9297859)
rs9329280  ~  rs7898275+rs12261462+rs3132006+rs7081782 (MAF = 0.05025126, R-squared = 0.
rs4880517  ~  rs7898275+rs9419498 (MAF = 0.04060914, R-squared = 0.9999999)
```

The rules are also selectable by SNP name for detailed examination:

```
> rules[c('rs11253563', 'rs2379080')]

rs11253563  ~  rs7093061+rs9419496+rs7898275+rs2246654 (MAF = 0.251269, R-squared = 0.97
rs2379080  ~  rs4880983+rs4880750+rs2379078+rs10794885 (MAF = 0.2085427, R-squared = 0.8
```

Rules are shown with a + symbol separating predictor SNPs. (It is important to know which SNPs were used for each imputation when checking imputed test results for artifacts.)

A summary table of all the 14,251 rules is generated by

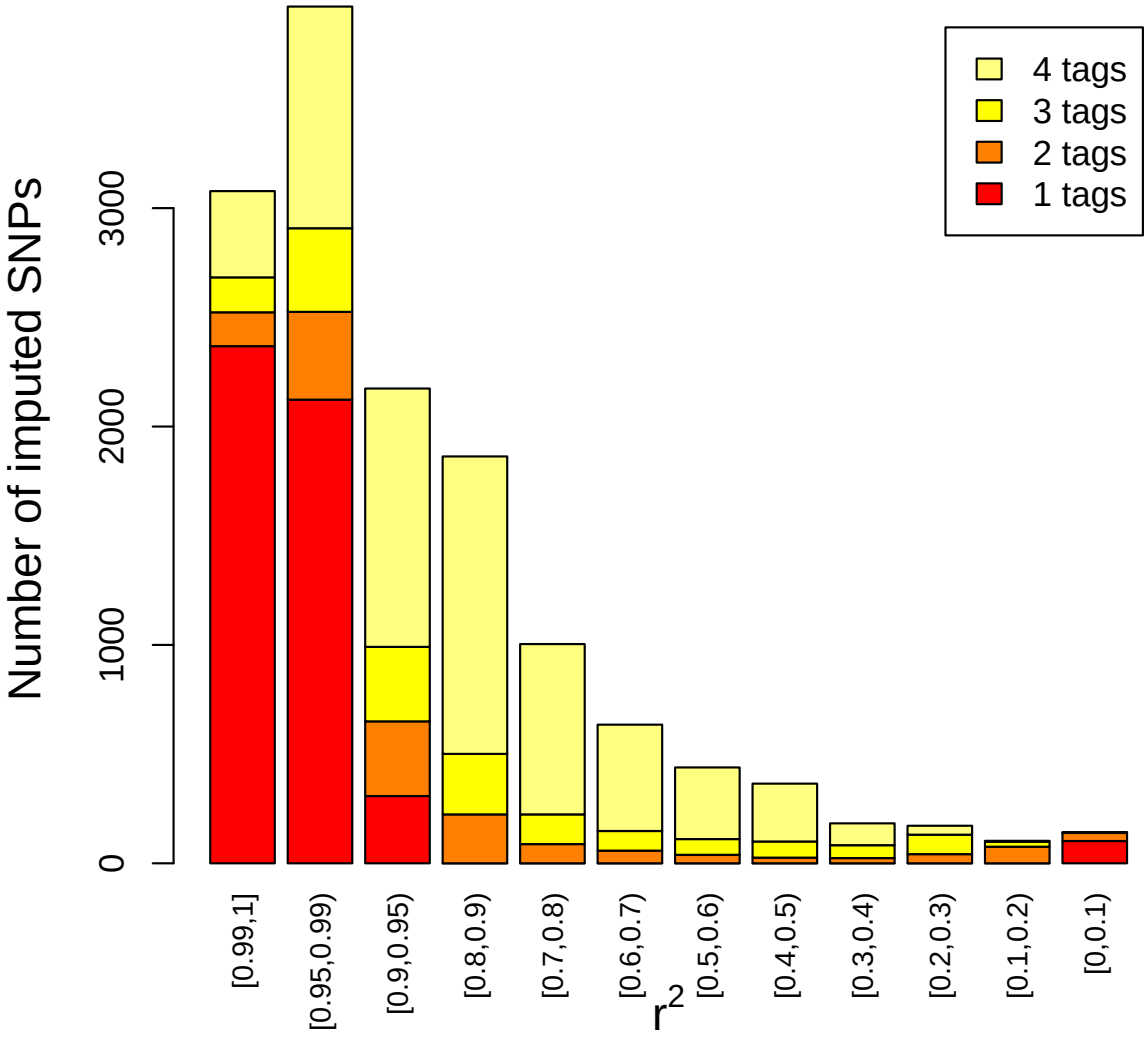
```
> summary(rules)
```

[2] For imputation from small samples, some smoothing of these haplotype frequencies would be advantageous and some ability to do this has been included. The `use.haps` argument to `snp.imputation` controls this. But invoking this option slows down the algorithm and it is not advised other than for very small sample sizes.

	SNPs used				
R-squared	1 tags	2 tags	3 tags	4 tags	<NA>
[0,0.1)	102	37	3	0	0
[0.1,0.2)	0	76	23	4	0
[0.2,0.3)	0	42	89	41	0
[0.3,0.4)	0	24	59	100	0
[0.4,0.5)	0	26	74	265	0
[0.5,0.6)	0	39	72	328	0
[0.6,0.7)	0	58	90	487	0
[0.7,0.8)	0	88	136	780	0
[0.8,0.9)	0	224	277	1362	0
[0.9,0.95)	308	342	341	1183	0
[0.95,0.99)	2123	402	383	1015	0
[0.99,1]	2368	155	160	395	0
<NA>	0	0	0	0	170

Columns represent the number of tag SNPs while rows represent grouping on R^2. The last column (headed `<NA>`) represents SNPs for which an imputation rule could not be computed, either because they were monomorphic or because there was insufficient data (as determined by the `minA` optional argument in the call to `snp.imputation`). The same information may be displayed graphically by

```
> plot(rules)
```



Carrying out the association tests

The association tests for imputed SNPs can be carried out using the function `single.snp.tests`.

```
> imp <- single.snp.tests(cc, stratum, data=subject.support,
+                         snp.data=target, rules=rules)
```

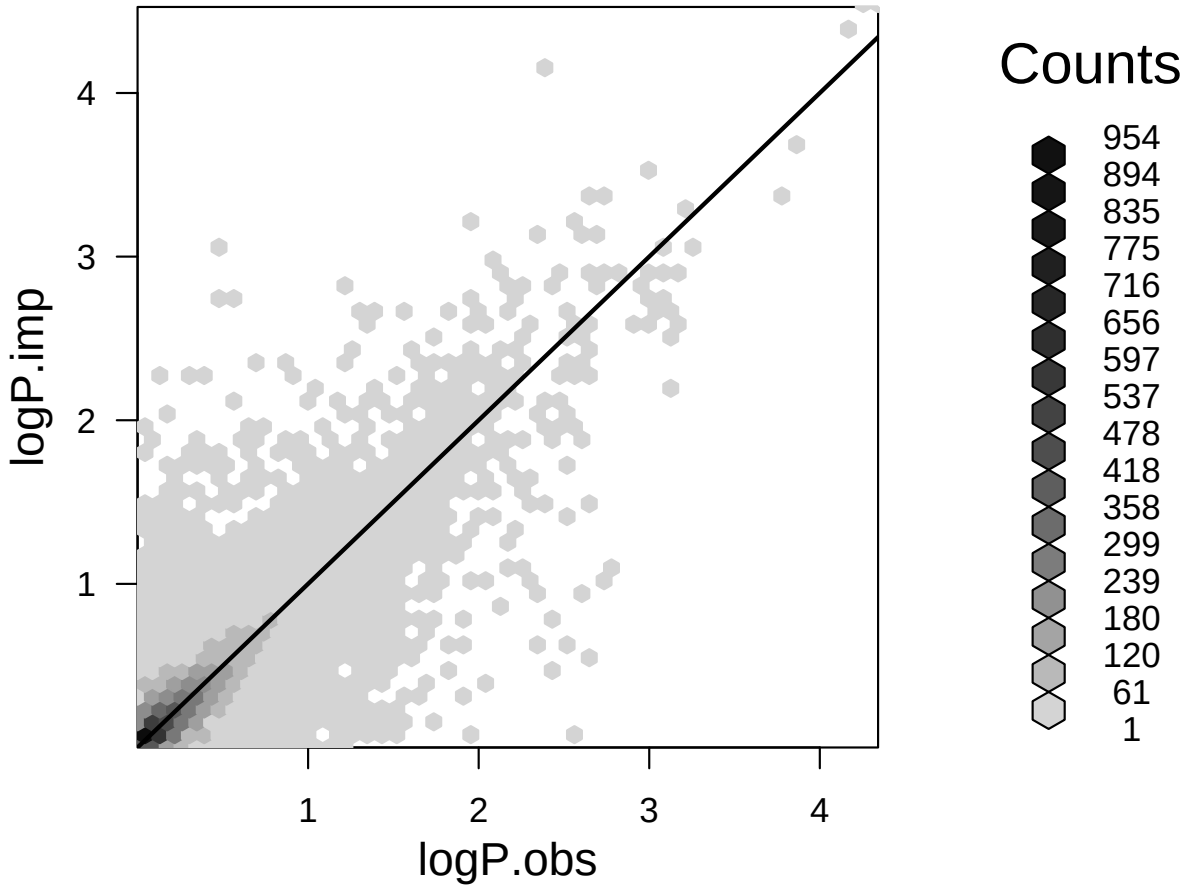
Using the observed data in the matrix `target` and the set of imputation rules stored in `rules`, the above command imputes each of the imputed SNPs, carries out 1- and 2-df

single locus tests for association, returns the results in the object `imp`. To see how successful imputation has been, we can carry out the same tests using the *true* data in `missing`:

```
> obs <- single.snp.tests(cc, stratum, data=subject.support, snp.data=lost)
```

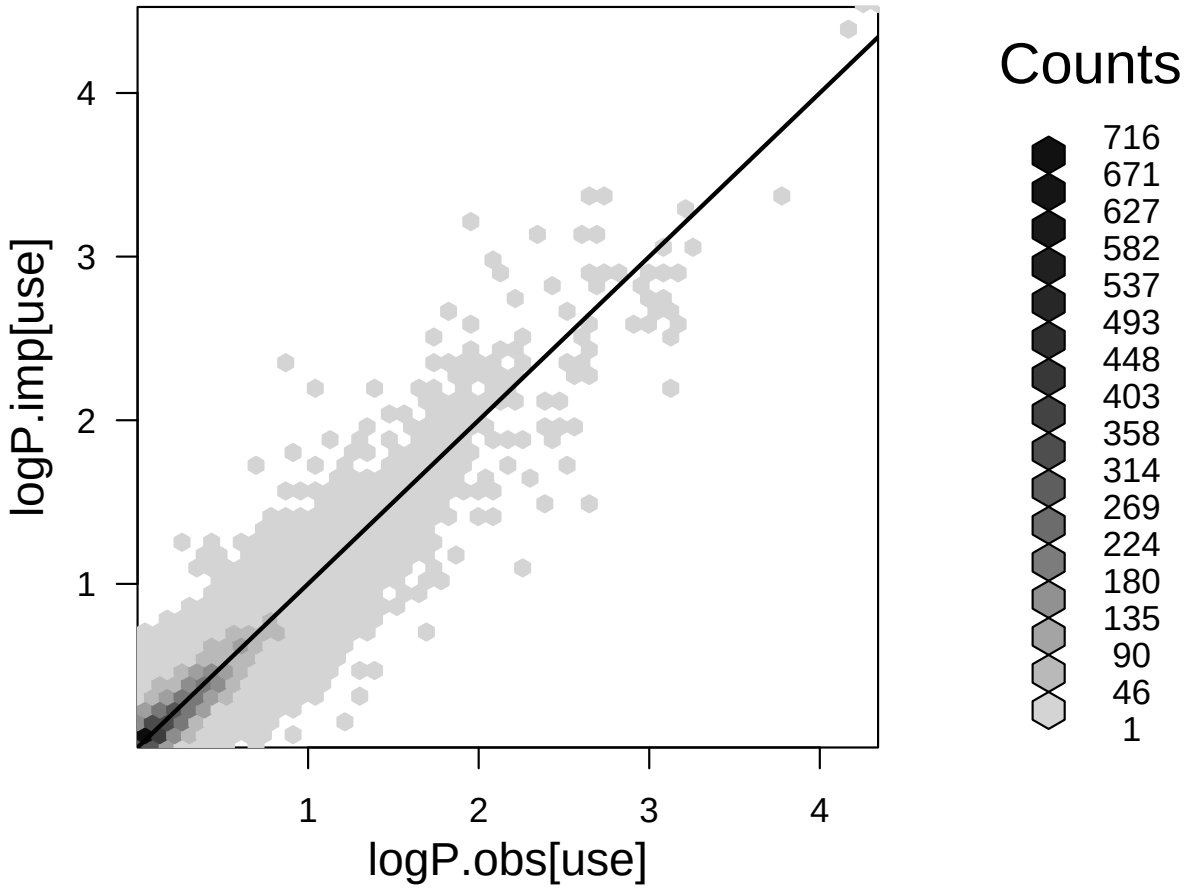
The next commands extract the p-values for the 1-df tests, using both the imputed and the true "missing" data, and plot one against the other (using the `hexbin` plotting package for clarity):

```
> logP.imp <- -log10(p.value(imp, df=1))
> logP.obs <- -log10(p.value(obs, df=1))
> hb <- hexbin(logP.obs, logP.imp, xbin=50)
> sp <- plot(hb)
> hexVP.abline(sp$plot.vp, 0, 1, col="black")
```



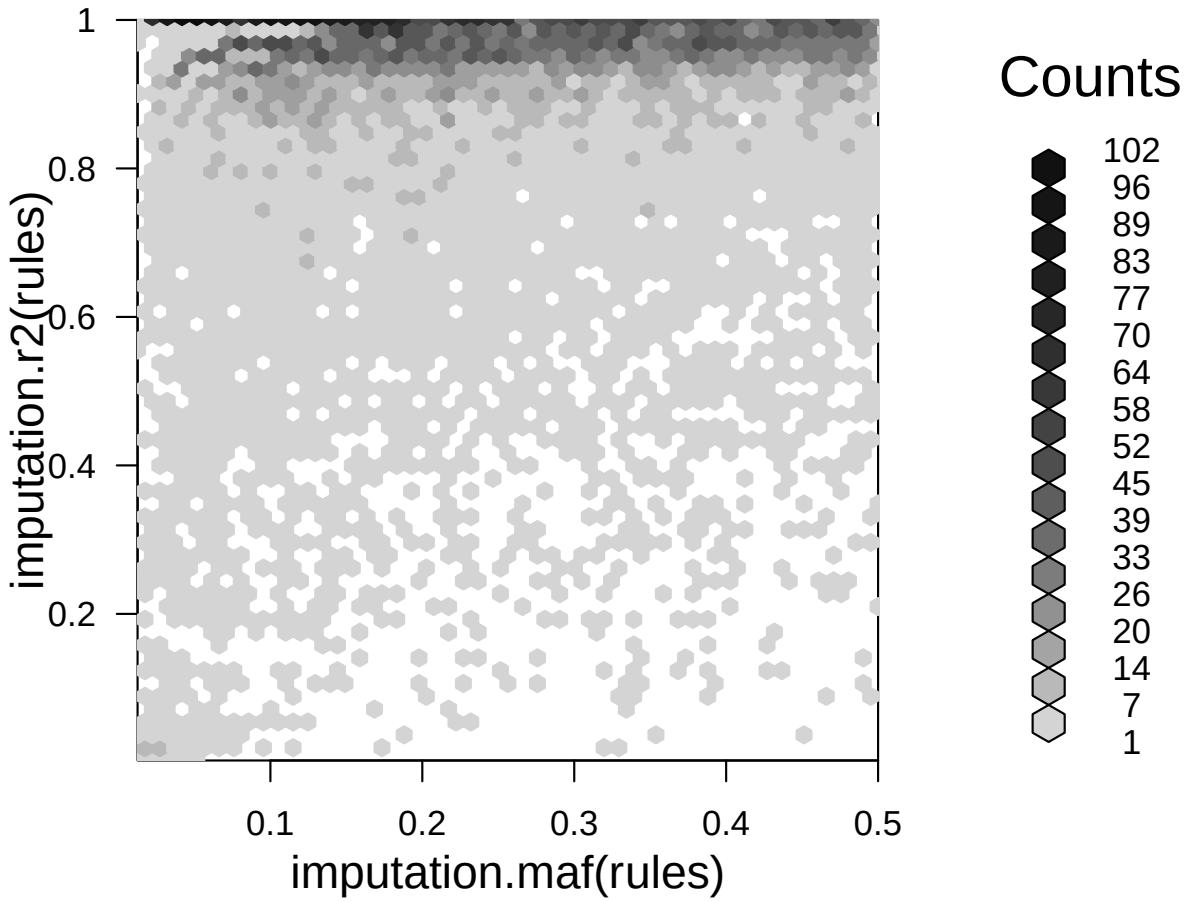
As might be expected, the agreement is rather better if we only compare the results for SNPs that can be computed with high R^2. The R^2 value is extracted from the `rules` object, using the function `imputation.r2` and used to select a subset of rules:

```
> use <- imputation.r2(rules)>0.9
> hb <- hexbin(logP.obs[use], logP.imp[use], xbin=50)
> sp <- plot(hb)
> hexVP.abline(sp$plot.vp, 0, 1, col="black")
```



Similarly, the function `imputation.maf` can be used to extract the minor allele frequencies of the imputed SNP from the `rules` object. Note that there is a tendency for SNPs with a high minor allele frequency to be imputed rather more successfully:

```
> hb <- hexbin(imputation.maf(rules), imputation.r2(rules), xbin=50)
> sp <- plot(hb)
```

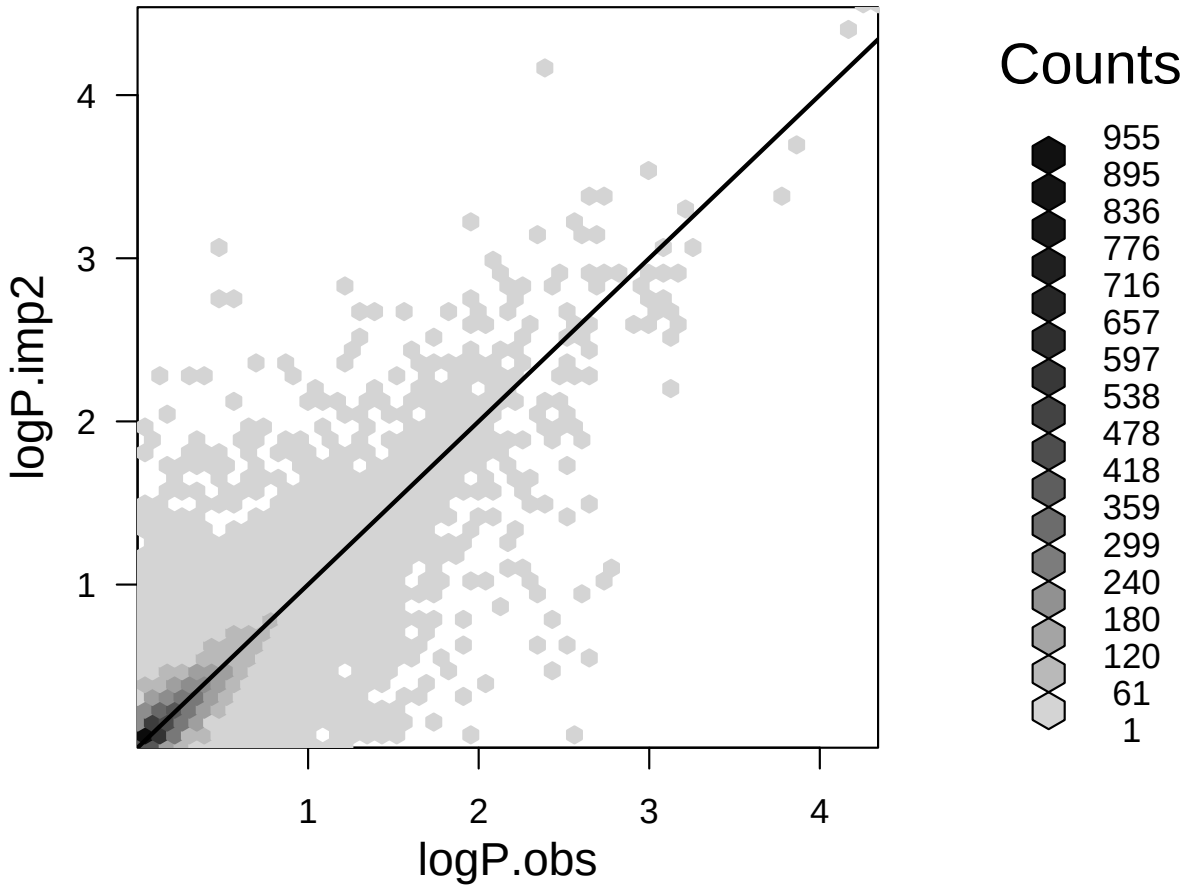


The function `snp.rhs.glm` also allows testing imputed SNPs. In its simplest form, it can be used to calculate essentially the same tests as carried out with `single.snp.tests`[3] (although, being a more flexible function, this will run somewhat slower). The next commands recalculate the 1 df tests for the imputed SNPs using `snp.rhs.tests`, and plot the results against those obtained when values are observed.

```
> imp2 <- snp.rhs.tests(cc~strata(stratum), family="binomial",
+                       data=subject.support, snp.data=target, rules=rules)
> logP.imp2 <- -log10(p.value(imp2))
```

[3]There is a small discrepancy, of the order of $(N-1) : N$.

```
> hb <- hexbin(logP.obs, logP.imp2, xbin=50)
> sp <- plot(hb)
> hexVP.abline(sp$plot.vp, 0, 1, col="black")
```



Storing imputed genotypes

In the previous two sections we have seen how to (a) generate imputation rules and, (b) carry out tests on SNPs imputed according to these rules, but without storing the imputed genotypes. It is also possible to store imputed SNPs in an object of class `SnpMatrix` (or

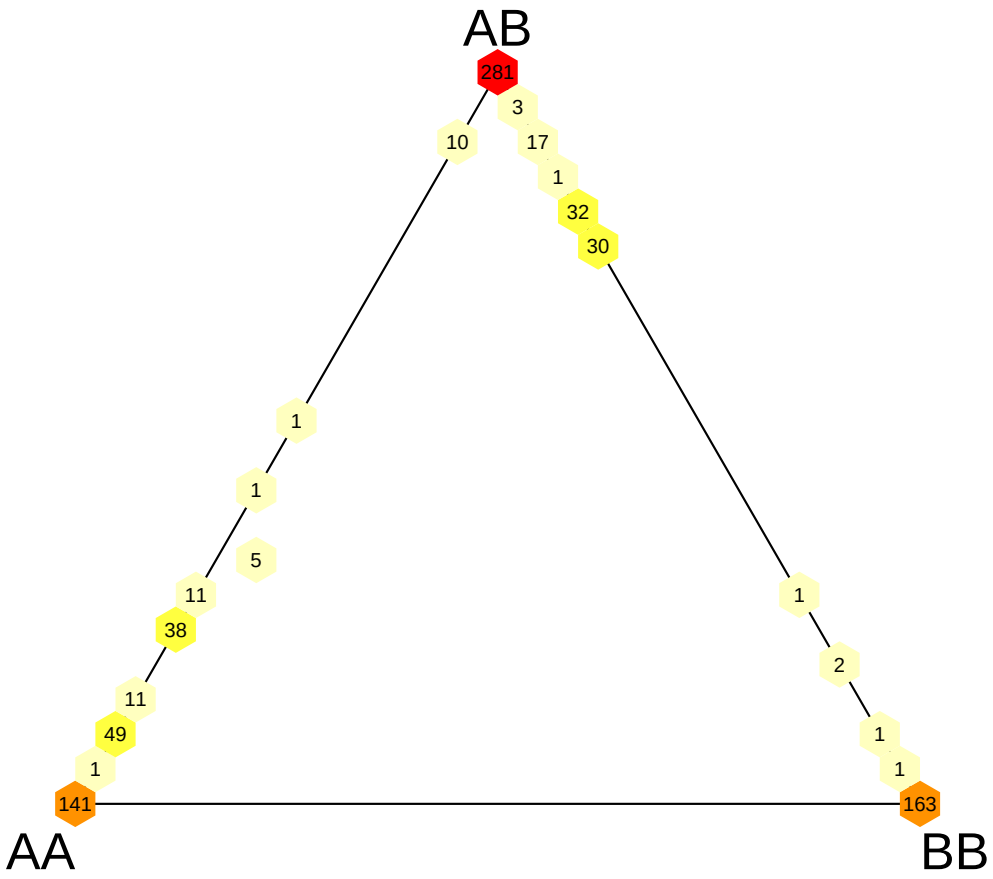
XSnpMatrix). The posterior probabilities of assignment of each individual to the three possible genotypes are stored within a one byte variable, although obviously not to full accuracy.

The following command imputes the "missing" SNPs using the "target" dataset and stores the imputed values in an object of class SnpMatrix:

```
> imputed <- impute.snps(rules, target, as.numeric=FALSE)
```

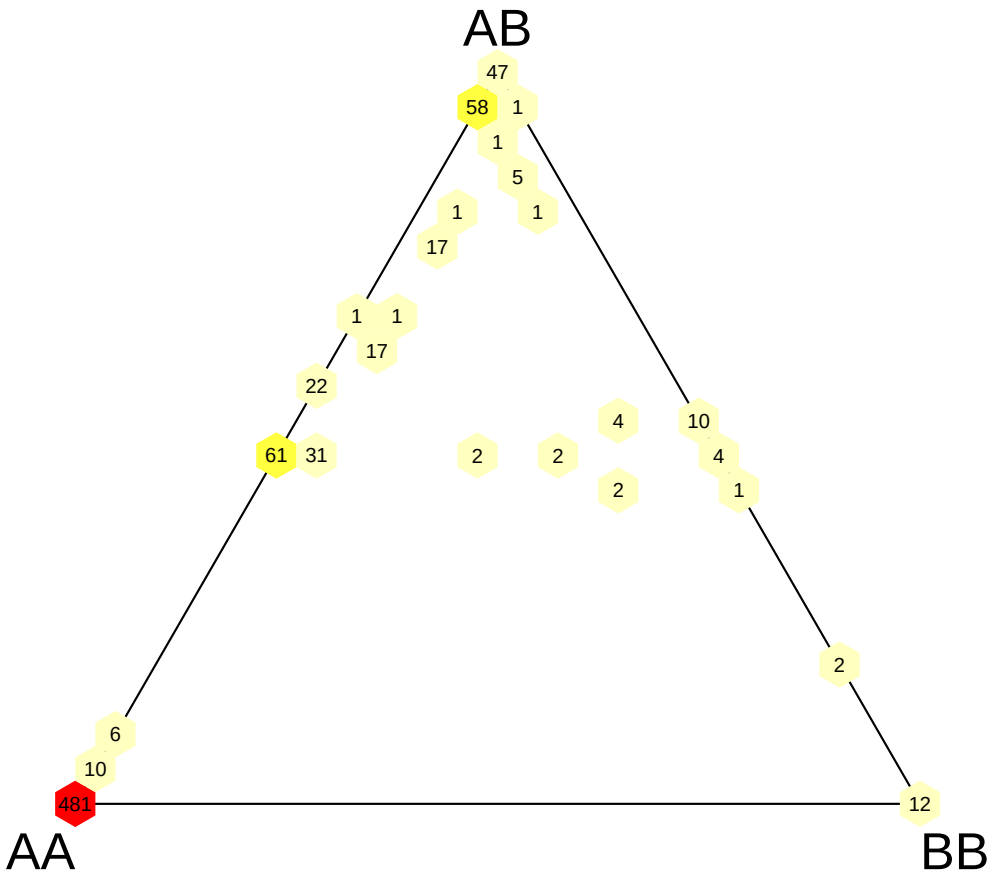
(If as.numeric were set to TRUE, the default, the resulting object would be a simple numeric matrix containing posterior expectations of the 0, 1, 2 genotype.) A nice graphical description of how snpStats stores uncertain genotypes is provided by the function plotUncertainty. This plots the frequency of the stored posterior probabilities on an equilateral triangle. The posterior probabilities are represented by the perpendicular distances from each side, the vertices of the triangle corresponding to certain assignments. Thus, the SNP rs4880568 is accurately imputed ($R^2 = 0.94$)

```
> plotUncertainty(imputed[, "rs4880568"])
```



while `rs2050968` is rather less so ($R^2 = 0.77$

```
> plotUncertainty(imputed[, "rs2050968"])
```



Tests can be carried out on these uncertainly assigned genotypes. For example

```
> imp3 <- single.snp.tests(cc, stratum, data=subject.support,
+                          snp.data=imputed, uncertain=TRUE)
```

The `uncertain=TRUE` argument ensures that uncertaing genotypes are used in the computations. This should yield nearly the same result as before. For the first five SNPs we have

```
> imp3[1:5]
```

```
             N Chi.squared.1.df Chi.squared.2.df     P.1df     P.2df
rs7909677  800       0.06279194        0.4887137 0.8021364 0.7832081
```

```
rs12773042 800       0.05676702        0.4873555 0.8116808 0.7837402
rs11253563 800       1.75508393        1.7577666 0.1852389 0.4152464
rs4881552  800       0.24752562        0.2643539 0.6188228 0.8761859
rs10904596 800       1.80595373        1.8176006 0.1789944 0.4030074
```

```
> imp[1:5]
```

```
             N     N.r2 Chi.squared.1.df Chi.squared.2.df     P.1df     P.2df
rs7909677  800 694.4234       0.07362092        0.5021694 0.7861359 0.7779565
rs12773042 800 731.0508       0.06371577        0.4861192 0.8007166 0.7842248
rs11253563 800 781.7304       1.74359665        1.7459346 0.1866838 0.4177102
rs4881552  800 779.0491       0.23668030        0.2527537 0.6266141 0.8812827
rs10904596 800 799.7839       1.79270628        1.8021045 0.1805968 0.4061421
```

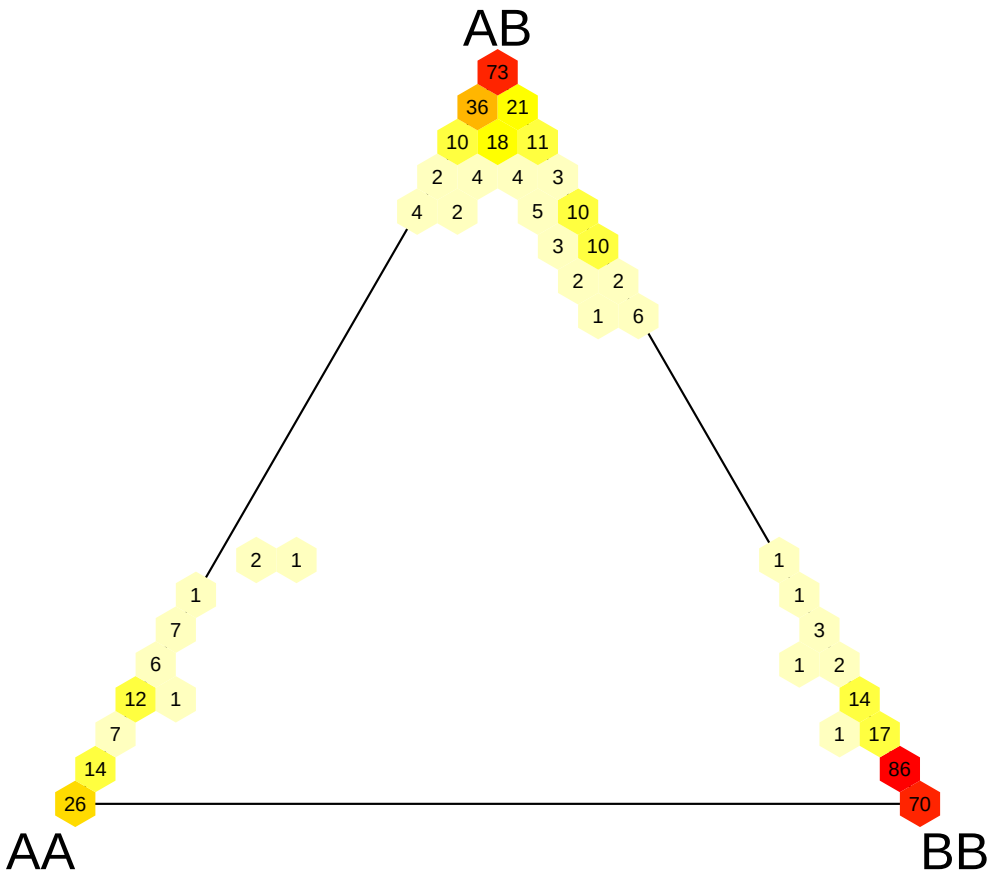
There are small discrepancies due to the genotype assignment probabilities not being stored to full accuracy. However these should have little effect on power of the tests and no effect on the type 1 error rate.

Note that the ability of `snpStats` to store imputed genotypes in this way allows alternative programs to be used to generate the imputed genotypes. For example, the file "mach1.out.mlprob.gz" (which is stored in the `extdata` sub-directory of the `snpStats` package) contains imputed SNPs generated by the MACH program, using the `-mle` and `-mldetails` options. In the following commands, we find the full path to this file, read it, and inspect one the imputed SNP in column 50:

```
> path <- system.file("extdata/mach1.out.mlprob.gz", package="snpStats")
> mach <- read.mach(path)
```

```
Reading MACH data from file /private/var/folders/r0/l4fjk6cj5xj0j3brt4bplpl40000gt/T/Rtm
Reading SnpMatrix with 500 rows and 178 columns
```

```
> plotUncertainty(mach[,50])
```



Meta-analysis

As stated at the beginning of this document, one of the main reasons that we need imputation is to perform meta-analyses which bring together data from genome-wide studies which use different platforms. The `snpStats` package includes a number of tools to facilitate this. All the tests implemented in `snpStats` are "score" tests. In the 1 df case we calculate a score defined by the first derivative of the log likelihood function with respect to the association parameter of interest at the parameter value corresponding to the null hypothesis of no

association. Denote this by U. We also calculate an estimate of its variance, also under the null hypothesis — V say. Then U^2/V provides the chi-squared test on 1 df. This procedure extends easily to meta-analysis; given two independent studies of the same hypothesis, we simply add together the two values of U and the two values of V, and then calculate U^2/V as before. These ideas also extend naturally to tests of several parameters (2 or more df tests).

In `snpStats`, the statistical testing functions can be called with the option `score=TRUE`, causing an extended object to be saved. The extended object contains the U and V values, thus allowing later combination of the evidence from different studies. We shall first see what sort of object we have calculated previously using `single.snp.tests` *without* the `score=TRUE` argument.

```
> class(imp)

[1] "SingleSnpTests"
attr(,"package")
[1] "snpStats"
```

This object contains the imputed SNP tests in our target set. However, these SNPs were observed in our training set, so we can test them. We will also recalculate the imputed tests. In both cases we will save the score information:

```
> obs <- single.snp.tests(cc, stratum, data=subject.support, snp.data=missing,
+                         score=TRUE)
> imp <- single.snp.tests(cc, stratum, data=subject.support,
+                         snp.data=target, rules=rules, score=TRUE)
```

The extended objects have been returned:

```
> class(obs)

[1] "SingleSnpTestsScore"
attr(,"package")
[1] "snpStats"

> class(imp)

[1] "SingleSnpTestsScore"
attr(,"package")
[1] "snpStats"
```

These extended objects behave in the same way as the original objects, so that the same functions can be used to extract chi-squared values, p-values etc., but several additional functions, or methods, are now available. Chief amongst these is `pool`, which combines evidence across independent studies as described at the beginning of this section. Although `obs` and `imp` are *not* from independent studies, so that the resulting test would not be valid, we can use them to demonstrate this:

```
> both <- pool(obs, imp)
> class(both)

[1] "SingleSnpTests"
attr(,"package")
[1] "snpStats"

> both[1:5]

             N     N.r2 Chi.squared.1.df Chi.squared.2.df     P.1df     P.2df
rs7909677  998 892.4234       0.04986583        0.5000542 0.8232969 0.7787797
rs12773042 997 928.0508       0.10603102        0.5084840 0.7447088 0.7755041
rs11253563 997 978.7304       1.76388425        2.1065032 0.1841408 0.3488017
rs4881552  997 976.0491       0.31648901        0.3915229 0.5737253 0.8222084
rs10904596 998 997.7839       1.75882469        1.7595965 0.1847712 0.4148666
```

Note that if we wished at some later stage to combine the results in `both` with a further study, we would also need to specify `score=TRUE` in the call to `pool`:

```
> both <- pool(obs, imp, score=TRUE)
> class(both)

[1] "SingleSnpTestsScore"
attr(,"package")
[1] "snpStats"
```

Another reason to save the score statistics is that this allows us to investigate the *direction* of findings. These can be extracted from the extended objects using the function `effect.sign`. For example, this command tabulates the signs of the associations in `obs`:

```
> table(effect.sign(obs))

  -1    0    1
7091   20 7140
```

In this table, -1 corresponds to tests in which effect sizes were negative (corresponding to an odds ratio less than one), while +1 indicates positive effect sizes (odds ratio greater than one). Zero sign indicates that the effect was `NA` (for example because the SNP was monomorphic). Reversal of sign can be the explanation of a puzzling phenomenon when two studies give significant results individually, but no significant association when pooled. Although it is not impossible that such results are genuine, a more usual explanation is that the two alleles have been coded differently in the two studies: allele 1 in the first study is allele 2 in the second study and vice versa. To allow for this, `snpStats` provides the `switch.alleles` function, which reverses the coding of specified SNPs. It can be applied to `SnpMatrix` objects but, because allele switches are often discovered quite late on in the

analysis and recoding the original data matrices could have unforeseen consequences, the `switch.alleles` function can also be applied to the extended test output objects. This modifies the saved scores *as if* the allele coding had been switched in the original data. The use of this is demonstrated below.

```
> effect.sign(obs)[1:6]

 rs7909677 rs12773042 rs11253563  rs4881552 rs10904596  rs4880781
        -1          -1           1          -1          -1           1

> sw.obs <- switch.alleles(obs, 1:3)
> class(sw.obs)

[1] "SingleSnpTestsScore"
attr(,"package")
[1] "snpStats"

> effect.sign(sw.obs)[1:6]

 rs7909677 rs12773042 rs11253563  rs4881552 rs10904596  rs4880781
         1           1          -1          -1          -1           1
```