Applied Statistical Genetics in R

for population-based association studies

PART III: Methods for ambiguity in haplotypic phase

Outline

1. Estimating haplotype frequencies
   - Expectation-maximization (EM) approach
   - Bayesian haplotype reconstruction
2. Estimating and testing haplotype-trait association
   - Two-staged approaches
   - General Linear Model (GLM) - EM approach

Estimating haplotype frequencies

- EM approach sets out to estimate haplotype frequences. In turn, these estimates can be used to infer unknown haplotypes for the individuals in our sample.
- Bayesian approach that focuses on reconstructing unknown haplotypes. In turn, the reconstructed data can be used to estimate population level haplotype frequencies.
- Reconstructed haplotypes can be used in regression modeling framework to characterize association (2-stage approach.)

Expectation-maximization (EM) approach

- Recall that a maximum likelihood estimate (MLE) is an estimate of a population level parameter, θ, that is derived by maximizing a function (the likelihood) of the data, $\mathbf{X} = (x_1, \ldots, x_n)$, given by:

$$L(\theta|\mathbf{X}) = \prod_{i=1}^{n} Pr(x_i|\theta)$$

  where $Pr(x_i|\theta)$ is the probability density function of x_i.
- In missing data settings, the data X are not fully observed and so this can not be calculated.
- The set of data that includes both the observed data, denoted $\mathbf{X}^{obs}$ and the missing data is commonly referred to as the *complete* data and is denoted $\mathbf{X}^{c}$.

Expectation-maximization (EM) approach

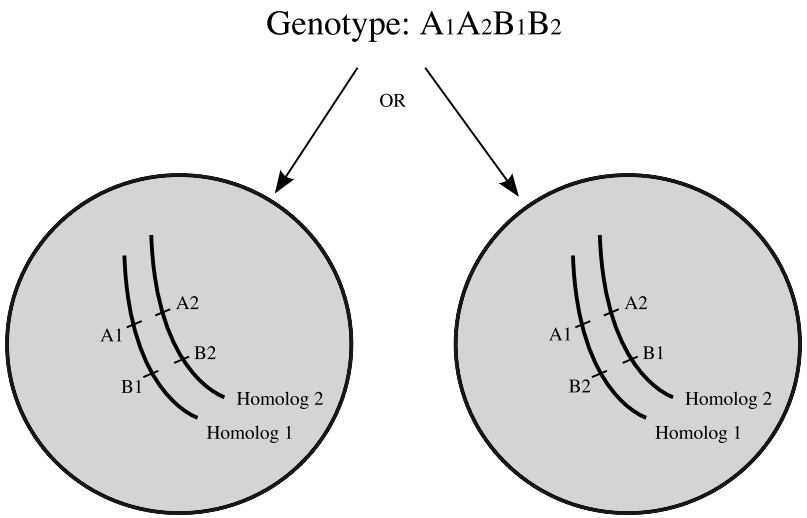
- *E-step*: Take the expectation of the complete data log likelihood conditional on the observed data and the current parameter estimate. Formally, we calculate:

$$E(l_\theta) = E\left(\log L(\theta|\mathbf{X}^c)|\mathbf{X}^{obs}, \widehat{\theta}^{(t)}\right)$$

  where $E(\cdot)$ is the expectation.
- *M-step*: Maximize $E(l_\theta)$ with respect to the parameter θ. This yields a new estimate, which we denote $\theta^{(t+1)}$.
- Iterate between the E-step and the M-step until a convergence criterion is met to arrive at a MLE of θ.

Expectation-maximization (EM) approach

- The observed data are the two nucleotides at each site and the missing data are the specific alignment of these nucleotides on each of the two homologous chromosomes.
  - Observed data = genotypes for all individuals in our sample
  - Missing data = corresponding haplotype pairs.
- The complete data are thus comprised of both the observed genotype information and the haplotypes pairs.



Expectation-maximization (EM) approach

- Derive complete data likelihood for the haplotypes
- Take conditional expectation of this function, conditional on the observed genotypes. This expectation is a weighted sum over all haplotype pairs that are consistent with the observed genotypes for a given an individual, where the weights are the *posterior probabilities* of the corresponding haplotype pair, given by:

$$Pr(H_i|G_i,\widehat{\theta}^{(t)}) = \frac{Pr(H_i;G_i|\widehat{\theta}^{(t)})}{Pr(G_i|\widehat{\theta}^{(t)})} = \frac{Pr(H_i|\widehat{\theta}^{(t)})}{\sum_H Pr(H_i;G_i|\widehat{\theta}^{(t)})}$$

where G_i in the genotype and H_i is the haplotype pair for individual i where $i = 1,\ldots,n$.

Expectation-maximization (EM) approach

- The expectation is then maximized to arrive at updated parameter estimates and the process is repeated.
- Importantly, this approach assumes HWE and thus should be applied within racial/ethnic strata within which there is no evidence of a departure from this assumption.

Expectation-maximization (EM) approach

- **Example 1** (EM approach to haplotype frequency estimation):
  - In this example we estimate the population level frequencies of haplotypes within the actn3 gene for African Americans and Caucasians separately based on the FAMuSS data.
  - The genotype matrix has a pair of adjacent columns for each SNP such that each column corresponds to one of the two observed alleles at the corresponding site. The order of the columns is assumed to correspond to the order of the sites on the chromosome.

```
> install.packages("haplo.stats")
> library(haplo.stats)
> Geno <- cbind(substr(actn3_r577x,1,1),substr(actn3_r577x,2,2),
+  substr(actn3_rs540874,1,1), substr(actn3_rs540874,2,2),
+  substr(actn3_rs1815739,1,1),substr(actn3_rs1815739,2,2),
+  substr(actn3_1671064,1,1),substr(actn3_1671064,2,2))
> SNPnames <- c("actn3_r577x","actn3_rs540874","actn3_rs1815739",
+ "actn3_1671064")
```

Expectation-maximization (EM) approach

- (example continued)
  - We then subset African Americans and Caucasians and apply the `haplo.em()` function to each group:

```
> Geno.C <- Geno[Race=="Caucasian" & !is.na(Race),]
> HaploEM <- haplo.em(Geno.C,locus.label=SNPnames)
> HaploEM
```

Expectation-maximization (EM) approach

- (example continued)

```
=============================================================================
                                  Haplotypes
=============================================================================
   actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064 hap.freq
1            C              A               C             G  0.00262
2            C              A               T             A  0.00935
3            C              A               T             G  0.01354
4            C              G               C             A  0.47292
5            C              G               C             G  0.01059
6            T              A               C             A  0.00066
7            T              A               T             A  0.00000
8            T              A               T             G  0.39890
9            T              G               C             A  0.08556
10           T              G               C             G  0.00000
11           T              G               T             A  0.00066
12           T              G               T             G  0.00520
=============================================================================
                                   Details
=============================================================================
lnlike =  -1285.389
lr stat for no LD =  2988.268 , df =  6 , p-val =  0
```

Expectation-maximization (EM) approach

- (example continued)

```
> Geno.AA <- Geno[Race=="African Am" & !is.na(Race),]
> HaploEM2 <- haplo.em(Geno.AA,locus.label=SNPnames)
> HaploEM2

=============================================================================
                                Haplotypes
=============================================================================
  actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064 hap.freq
1           C              A               C             A  0.01053
2           C              A               C             G  0.08131
3           C              A               T             G  0.03764
4           C              G               C             A  0.57763
5           C              G               C             G  0.01086
6           T              A               C             A  0.00119
7           T              A               T             G  0.17166
8           T              G               C             A  0.10833
9           T              G               C             G  0.00086
=============================================================================
                                  Details
=============================================================================
lnlike =  -84.97972
lr stat for no LD =  119.2917 , df =  4 , p-val =  0
```

Expectation-maximization (EM) approach

- (example continued)
  - The column entitled `hap.freq` is the estimated population level haplotype frequency.
  - Based on this output we can see that the most prevalent haplotype is the same in African Americans and Caucasians and is given by $h_4 = (CGCA)$.
  - The estimated prevalence of this haplotype is higher for African Americans ($\widehat{\theta}_4 = 0.58$) than Caucasians ($\widehat{\theta}_4 = 0.47$).
  - On the other hand, the estimate prevalence of the $(TATG)$ haplotype (labelled h_7 and h_8 in the above output) is markedly lower in African Americans ($\widehat{\theta}_7 = 0.17$) than in Caucasians ($\widehat{\theta}_8 = 0.40$).

Expectation-maximization (EM) approach

- Testing hypotheses that involve the haplotype frequencies within the EM context requires consideration of the uncertainty in the estimation procedures.
- Construct a confidence interval around the difference in the two frequencies and check whether it covers 0.
  1. First we calculate the difference in the estimated frequencies.
  2. In order to calculate the standard error of each frequency, we use the function `HapFreqSE()` defined below.
  3. We then combine the results to get an estimate of the standard error of this difference.
  4. Finally a 95% confidence interval is calculated based on a normal probability distribution

Expectation-maximization (EM) approach

```
#############################################################
# Description: This function creates a design matrix with i,j
# element equal to the conditional expectation
# of the number of copies of haplotype j for
# individual i based on the output from haplo.em()
# Input: HaploEM (object resulting from haplo.em()), Output: XmatHap
#############################################################
HapDesign <- function(HaploEM){
Nobs <- length(unique(HaploEM$indx.subj)) # number of observations
Nhap <- length(HaploEM$hap.prob) # number of haplotypes
XmatHap <- matrix(data=0,nrow=Nobs,ncol=Nhap)
for (i in 1:Nobs){
IDSeq <- seq(1:sum(HaploEM$nreps))[HaploEM$indx.subj==i]
for (j in 1:length(IDSeq)){
XmatHap[i,HaploEM$hap1code[IDSeq][j]] <-
XmatHap[i,HaploEM$hap1code[IDSeq][j]] +
HaploEM$post[IDSeq][j]
XmatHap[i,HaploEM$hap2code[IDSeq][j]] <-
XmatHap[i,HaploEM$hap2code[IDSeq][j]] +
HaploEM$post[IDSeq][j]
}
}
return(XmatHap)
}
```

Expectation-maximization (EM) approach

```
#############################################################
# Description: This function creates a vector with jth element
# equal to the standard error of haplotype j
# based on the output from haplo.em()
# Input:  HaploEM (object resulting from haplo.em()), Output:  HapSE
#############################################################
HapFreqSE <- function(HaploEM){
HapMat <- HapDesign(HaploEM)
Nobs <- length(unique(HaploEM$indx.subj)) # number of observations
Nhap <- length(HaploEM$hap.prob) # number of haplotypes
S.Full<-matrix(data=0, nrow=Nobs, ncol=Nhap-1)
for(i in 1:Nobs){
for(k in 1:(Nhap-1)){
S.Full[i,k]<-HapMat[i,k]/HaploEM$hap.prob[k]-
HapMat[i,Nhap]/HaploEM$hap.prob[Nhap]
}
}
Score<-t(S.Full)%*%S.Full
invScore<-solve(Score)
HapSE<-c(sqrt(diag(invScore)),
sqrt(t(rep(1,Nhap-1))%*%invScore%*%rep(1,Nhap-1)))
return(HapSE)
}
```

Expectation-maximization (EM) approach

- **Example 2**: (Testing hypotheses about haplotype frequencies within the EM framework)
  - In this example we test the null hypothesis that the frequency of haplotype h_4 defined above is the same in African Americans and Caucasians.

```
> FreqDiff <- HaploEM2$hap.prob[4] - HaploEM$hap.prob[4]
> s1 <- HapFreqSE(HaploEM)[4]
> s2 <- HapFreqSE(HaploEM2)[4]
> SE <- sqrt(s1^2 + s2^2)
> CI <- c(FreqDiff - 1.96*SE, FreqDiff + 1.96*SE)
> CI
```

```
[1] -0.1538784  0.3632976
```

  - Since this interval covers 0 there is not enough evidence to suggest a difference in the frequency of h_4 between Caucasians and African Americans based on our sample and a 0.05 level test.

Expectation-maximization (EM) approach

- (example continued)
  - Based on the estimated population level haplotype probabilities, we can calculate the probabilities of each possible haplotype pair for an observation in our sample.
  - Example: (Calculating posterior haplotype probabilities) In this example we illustrate how to determine the posterior probabilities of each haplotype pair that is consistent with the observed genotype for an individual.
  - Recall `HapoEM` is the result of applying the `haplo.em()` function to the SNPs within the `actn3` gene on the Caucasian subgroup within the FAMuSS study.
  - The associated object `HaploEM$nreps` is a vector of length equal to the number of individuals in our sample with elements equal to the number of haplotype pairs that are consistent with the observed genotype. For example, consider the first 5 elements of this vector, given by:

Expectation-maximization (EM) approach

- (example continued)

```
> HaploEM$nreps[1:5]
```

```
indx.subj
1 2 3 4 5
1 3 3 3 1
```

  - This tells us that there is one haplotype pair consistent with the observed genotype for the first and fifth individuals and three pairs that are consistent with each of the observed genotypes for the second, third and fourth individuals.
  - The corresponding potential haplotypes for these five individuals are given by the associated vectors `HaploEM$hap1code` and `HaploEM$hap2code` as shown below where the coding corresponds to the numbering system we saw for Caucasians:

```
> HaploEM$hap1code[1:11]
```

```
 [1]  4  3 11  4  9 11  4  9  1  8  4
```

```
> HaploEM$hap2code[1:11]
```

```
 [1]  4  9  1  8  3  1  8  3 11  4  4
```

Expectation-maximization (EM) approach

- (example continued)
  - The `indx.subj` vector tells us the corresponding record number and contains the sequence of numbers from 1 to the number of observations in our sample, with each element of the sequence repeated according to the value in `HaploEM$nreps`.

```
> HaploEM$indx.subj[1:11]
 [1] 1 2 2 2 3 3 3 4 4 4 5
```

  - Based on this output, we see that the first and fifth individual have the haplotype pair $(4, 4)$ while the second, third and fourth individuals are all ambiguous between $(3, 9)$, $(11, 1)$ and $(4, 8)$.

Expectation-maximization (EM) approach

- (example continued)
  - The posterior probabilities associated with these pairs are given by:

```
> HaploEM$post[1:11]

 [1] 1.000000e+00 6.103655e-03 9.075404e-06 9.938873e-01 6.103655e-03
 [6] 9.075404e-06 9.938873e-01 6.103655e-03 9.075404e-06 9.938873e-01
[11] 1.000000e+00
```

  - Notably, the sum of these probabilities within any single individual is equal to 1.
  - We could have also calculate these probabilities directly based on the estimated haplotype frequencies. To see this, first note that the twelve haplotype probabilities are contained in the following vector:

```
> HapProb <- HaploEM$hap.prob
> HapProb

 [1] 2.618539e-03 9.346357e-03 1.353985e-02 4.729207e-01 1.059499e-02
 [6] 6.572523e-04 4.492090e-08 3.988971e-01 8.556339e-02 1.349443e-07
[11] 6.578380e-04 5.203819e-03
```

Expectation-maximization (EM) approach

- (example continued)
  - Now consider one of our individuals who is ambiguous between the pairs $(3, 9)$, $(11, 1)$ and $(4, 8)$. Assuming independence (which was already assumed in the estimation procedure), estimated probabilities of each of these pairs are given respectively by:

```
> p1 <- 2*prod(HapProb[c(3,9)])
> p2 <- 2*prod(HapProb[c(11,1)])
> p3 <- 2*prod(HapProb[c(4,8)])
> p1 / (p1+p2+p3)

[1] 0.006103654

> p2 / (p1+p2+p3)

[1] 9.075404e-06

> p3 / (p1+p2+p3)

[1] 0.9938873
```

  - As expected, these values are equivalent to the probabilities given in the second, third and fourth elements of `HaploEM$post`.

Expectation-maximization (EM) approach

- Analysts tend to assign individual the haplotype with the highest posterior probability and then treat these as known in subsequent analysis. (Single Imputation)
- Caution: Valuable information on the uncertainty in the assignment is lost.
- Instead, methods specifically developed for this setting can be applied if the ultimate goal is to characterize haplotype-trait association.

(Bayesian haplotype reconstruction)

Two-staged approaches to association

- Haplotype trend regression (HTR)
  - Assign to each ambiguous individual in our sample the *conditional expectation* of the number of copies of each potential haplotype. That is, in place of indicating the presence of $n_h = 0, 1$ or 2 copies of haplotype h, as we would do in the observed data setting, we instead assign a value of $E(n_h|G)$.
- **Example 3** (Application of haplotype trend regression (HTR)):
  1. Create a new design matrix with elements equal to the conditional expectation for the number of copies of each haplotype:

     ```
     > HapMat <- HapDesign(HaploEM)
     ```

Haplotype trend regression (HTR)

- (example continued)
  2. Fit a a linear model using this new design matrix and conduct an overall F-test which compares this model to the reduced model with just an intercept.

```
> Trait <- NDRM.CH[Race=="Caucasian" & !is.na(Race)]
> mod1 <- (lm(Trait~HapMat))
> mod2 <- (lm(Trait~1))
> anova(mod2,mod1)

Analysis of Variance Table

Model 1: Trait ~ 1
Model 2: Trait ~ HapMat
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    776 881666
2    764 868364  12     13303 0.9753 0.4708
```

- Based on this output we conclude that there is not sufficient evidence to support a haplotype-trait association.

(Multiple imputation)

GLM-EM approach

- Fully likelihood-based approach that involves simultaneous estimation of haplotype frequencies and measures of association.
- Advantage of this approach is that it uses information on the trait to inform us about the most likely haplotpye pair for an individual.
- Involves the application of an expectation-maximization (EM) algorithm. In this setting, however, the complete data log likelihood is now defined in terms of both the haplotypes and the trait under investigation.

GLM-EM Approach

- **Example 4** (EM for estimation and testing of haplotype-trait association):
  - Again consider data from the FAMuSS study and association between haplotypes within the `actn3` gene and the percent change in the non-dominant arm muscle strength.
  - Using the genotype data matrix `Geno.C` that we generated above we can apply the following code, where again `haplo.glm()` is a function within the `haplo.stats` package:

```
> install.packages("haplo.stats")
> library(haplo.stats)
> Geno.C <- setupGeno(Geno.C)
> Dat <- data.frame(Geno.C=Geno.C, Trait=Trait)
> haplo.glm(Trait~Geno.C,data=Dat,
+ allele.lev=attributes(Geno.C)$unique.alleles)
```

GLM-EM Approach

- (example continued)
  - This results in the following:

```
  Call:
haplo.glm(formula = Trait ~ Geno.C, data = Dat,
    allele.lev = attributes(Geno.C)$unique.alleles)

Coefficients:
              coef     se  t.stat   pval
(Intercept) 50.682  2.174 23.3181 0.0000
Geno.C.2    -7.646  8.423 -0.9077 0.3643
Geno.C.3     8.612  3.568  2.4134 0.0160
Geno.C.5    -0.398  7.250 -0.0549 0.9562
Geno.C.8     1.967  1.877  1.0482 0.2949
Geno.C.9     8.443  3.499  2.4129 0.0161
Geno.C.12   19.659 10.079  1.9505 0.0515
Geno.C.rare 11.859  0.176 67.4324 0.0000
```

GLM-EM Approach

- (example continued)

```
Haplotypes:
            loc.1 loc.2 loc.3 loc.4 hap.freq
Geno.C.2        C     A     T     A  0.00884
Geno.C.3        C     A     T     G  0.01248
Geno.C.5        C     G     C     G  0.01078
Geno.C.8        T     A     T     G  0.40244
Geno.C.9        T     G     C     A  0.08399
Geno.C.12       T     G     T     G  0.00528
Geno.C.rare     *     *     *     *  0.00401
haplo.base      C     G     C     A  0.47218
```

  - By default, the `base` haplotype is set equal to the most prevalent one in our sample, in this case $h_4 = (CGCA)$.
  - Each of the p-values returned by applying the `haplo.glm()` function correspond to a test that the effect of the corresponding haplotype, compared to this base haplotype, is equal to 0.

GLM-EM Approach

- (example continued)
  - For example, we see that the effect of h_9 = `Geno.C.9` is 8.443 with a corresponding p-value of 0.0161.
  - This implies that the mean percent change in muscle strength is 8.443 greater among individuals with one copy of the h_9 haplotype compared to individuals that are homozygous for the h_4 haplotype and this effect is significantly different than zero.
- Several parameters can be controlled in the above analysis.
- First of all, we can change the base haplotype, which is useful if we are trying to compare results across different racial/ethnic strata within which the most prevalent haplotype differs.
- This is also useful if there is specific haplotype to which we would like to compare all other haplotypes.

GLM-EM Approach

- **Example 5**
  - Suppose we want $h_9 = (TGCA)$ to be the base haplotype.

```
> haplo.glm(Trait~Geno.C,data=Dat,
+ allele.lev=attributes(Geno.C)$unique.alleles,
+ control=haplo.glm.control(haplo.base=9))

  Call:
haplo.glm(formula = Trait ~ Geno.C, data = Dat,
    allele.lev = attributes(Geno.C)$unique.alleles,
    control = haplo.glm.control(haplo.base = 9))

Coefficients:
               coef    se  t.stat     pval
(Intercept)  67.569 4.697 14.3852 0.00e+00
Geno.C.2    -16.089 8.052 -1.9982 4.60e-02
Geno.C.3      0.169 2.383  0.0708 9.44e-01
Geno.C.4     -8.443 2.820 -2.9940 2.84e-03
Geno.C.5     -8.841 7.174 -1.2323 2.18e-01
Geno.C.8     -6.476 2.766 -2.3410 1.95e-02
Geno.C.12    11.216 4.668  2.4028 1.65e-02
Geno.C.rare   3.416 0.641  5.3300 1.29e-07
```

GLM-EM Approach

- (example continued)

```
Haplotypes:
            loc.1 loc.2 loc.3 loc.4 hap.freq
Geno.C.2        C     A     T     A  0.00884
Geno.C.3        C     A     T     G  0.01248
Geno.C.4        C     G     C     A  0.47218
Geno.C.5        C     G     C     G  0.01078
Geno.C.8        T     A     T     G  0.40244
Geno.C.12       T     G     T     G  0.00528
Geno.C.rare     *     *     *     *  0.00401
haplo.base      T     G     C     A  0.08399
```

- Another important parameter that we can control is an indicator for the type of genetic model.
- By default, the `haplo.glm()` function assumes an additive genetic model, so that the effect of having two copies of a haplotype is twice the effect of having a single copy of the haplotype.

GLM-EM Approach

- We can easily specify an alternative model structure by specifying the `haplo.effect` parameter within `haplo.glm.control`.
  - For example, under a dominant genetic model in which one or more copies of a haplotype causes the effect, we have:

```
> haplo.glm(Trait~Geno.C,data=Dat,
+   allele.lev=attributes(Geno.C)$unique.alleles,
+   control=haplo.glm.control(haplo.effect="dominant"))
```

GLM-EM Approach

- (example continued)

```
Coefficients:
              coef     se t.stat    pval
(Intercept) 49.06  2.326 21.087 0.00000
Geno.C.2    -9.85 10.526 -0.936 0.34967
Geno.C.3     4.34  1.935  2.241 0.02531
Geno.C.5     1.38  9.394  0.147 0.88326
Geno.C.8     4.42  2.675  1.654 0.09847
Geno.C.9    13.86  4.854  2.855 0.00442
Geno.C.12   14.27  0.636 22.438 0.00000
Geno.C.rare 12.46  0.261 47.722 0.00000

Haplotypes:
            loc.1 loc.2 loc.3 loc.4 hap.freq
Geno.C.2        C     A     T     A  0.00882
Geno.C.3        C     A     T     G  0.01254
Geno.C.5        C     G     C     G  0.01078
Geno.C.8        T     A     T     G  0.40240
Geno.C.9        T     G     C     A  0.08383
Geno.C.12       T     G     T     G  0.00529
Geno.C.rare     *     *     *     *  0.00400
haplo.base      C     G     C     A  0.47234
```

GLM-EM Approach

- (example continued)
  - Results of this analysis are slightly different than the results under the additive assumption.
  - Specifically, there is stronger evidence of an association between haplotypes $h_3 = (CATG)$, $h_9 = (TGCA)$ and $h_{12} = (TGTG)$ and our trait.
- Other parameters that can be controlled include handling of missing genotype data, the minimum prevalence for inclusion in the `rare` haplotype group and the distributional family.
- The default setting handles missing genotype information within the EM framework while excluding individuals with missing trait or covariate data.
- The `family` option allows of testing association with a binary trait, similar to the `glm()`function.

GLM-EM Approach

- In general, the fully likelihood based approach described in this section is preferable to the two-staged approaches when the primary aim is characterize haplotype-trait association.
- The two-staged approaches, however, have have the advantage that they allow for estimation of haplotype effects within subgroups and then combining the data for estimation and testing of effects.
- For example, we can use the `haplo.em()` function within African Americans and Caucasians separately to arrive at posterior probability estimates of each individual and impute data based on these estimates. We are then able to fit a model using all of the data combined, giving us more power.
- The `haplo.glm()` function on the other hand makes the HWE assumption and is therefore most appropriately applied within racial/ethnic groups within which there is no evidence of a departure from HWE.

Summary

- Topics covered:
  - EM approach to haplotype frequency estimation
  - Haplotype trend regression
  - GLM-EM to test haplotype-trait association
- Useful R packages and functions:
  - haplo.stats: setupGeno(); haplo.em(); haplo.glm()
  - hapassoc: hapassoc()
  - haplo.ccs: haplo.ccs()
  - generic: substr()