Haplotype Estimation (Phasing)

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

November 21, 2018

Contents

Haplotype estimation

- Genetic marker data is often given at the genotype level (for diploid individuals).
- The specific alignment of alleles for markers on chromosomes (haplotypes) remains unknown.
- Statistical methods for inferring haplotypic phase have been proposed.

Why haplotype estimation?

- The number of markers used in genotyping studies is often very large. If we combine markers into haplotypes, the number of variables decreases. Association studies can be carried out between haplotypes and traits and reduce the problem of multiple testing.
- The purpose of haplotype estimation can be twofold:
  - resolving the haplotype constitution of each individual in the database.
  - estimation of population haplotype frequencies.
- Haplotypes may be biologically more relevant than single SNPs.
- Association studies with haplotypes tend to have more power than those using single markers.

Haplotypes

- A haplotype is a combination of alleles at adjacent loci on a chromosome that are transmitted together to the next generation.
- In practice, a haplotype often refers to a set of SNPs on a single chromosome that are statistically associated.
- In the previous module on LD we learned how to estimate haplotype frequencies by maximum likelihood.
- The EM algorithm can also be used to estimate haplotype frequencies and is implemented in the R package `haplo.stats`.
- The precise haplotypic composition of each individual in the database remains unknown, but can be estimated once the haplotype frequencies are known.

Example: two loci with two alleles

First locus has alleles A an a, second locus has alleles B and b

Genotypes	Corresponding haplotypes
AABB	AB and AB
AABb	AB and Ab
AAbb	Ab and Ab
AaBB	AB and aB
AaBb	AB and ab OR Ab and aB
Aabb	Ab and ab
aaBB	aB and aB
aaBb	aB and ab
aabb	ab and ab

Diplotypes

- The haplotypic constitution of an individiual is called its diplotype.
- E.g. if there are four haplotypes h_1, h_2, h_3 and h_4, the possible diplotypes are $(h_1, h_1), (h_1, h_2), (h_1, h_3), (h_1, h_4), (h_2, h_2), (h_2, h_3), (h_2, h_4), (h_3, h_3), (h_3, h_4)$ and (h_4, h_4).
- With two loci and two alleles (A, a and B, b), for the double heterozygote AaBb there are two possible diplotypes (AB,ab) and (Ab,aB).
- With three loci and two alleles (A,a), (B,b) and (C,c), for the triple heterozygote AaBbCc there are four possible diplotypes (ABC,abc), (ABc,abC), (AbC,aBc) and (aBC,Abc).
- For a multilocus genotype consisting of k heterozygous SNPs there are 2^{k-1} diplotypes.
- Note that for homozygous or single-site heterozygous individuals the haplotypes (and diplotypes) are known.

Methods for Haplotype estimation

- Parsimony methods. Resolve diplotypes with the minimum possible number of haplotypes.
- Likelihood based methods.
- Bayesian methods.
- ...

Parsimony methods

1. Identify all unambiguous haplotypes (homozygotes and single-site heterozygotes)
2. Check if any of the resolved haplotypes is compatible for an unresolved individual. If not, stop.
3. Identify the complementary haplotype for this last unresolved individual and add it to the set of resolved haplotypes. Return to (2)
4. Continue till all haplotypes are resolved, or no new haplotypes can be found.

Problems:

- There may be no unambiguous haplotypes.
- Unresolved haplotypes may remain.
- The solution is order-dependent.

Parsimony methods: toy example

	A/T	A/G	C/T	type
ID1	AT	AA	CT	(double heterozygote)
ID2	AT	GG	CT	(double heterozygote)
ID3	AA	AG	CC	(single-site heterozygote)
ID4	TT	GG	CT	(single-site heterozygote)
ID5	AA	AA	TT	(only homozygous)

- For ID3, ID4, and ID5 the haplotypes are unambiguous, and the resolved diplotypes are: (AAC,AGC), (TGC,TGT), (AAT,AAT)
- For ID1, AAC is compatible with the resolved haplotypes. Its complementary haplotype is TAT. Its diplotype may be inferred as (AAC,TAT). TAT is added to the pool of available haplotypes
- For ID2, AGC and TGT are both compatible with the resolved haplotypes. Its diplotype may be inferred as (AGC,TGT)
- Final haplotype set: (AAC,AGC,TGC,TGT,AAT,TAT)

Notes:

- Resolving ID2 as (AGT,TGC) would require creating a new haplotype
- If ID1 would be resolved as (AAT,TAC), the final haplotype set would be (AAC,AGC,TGC,TGT,AAT,TAC), which may be considered more plausible

Likelihood-based methods

- We have n individuals genotyped at m SNPs, yielding genotype data $\mathbf{G}$
- There are 2^m possible haplotypes with frequencies $\mathbf{h} = (h_1, \ldots h_{2^m})$

We wish to maximize

$$f(\mathbf{G}|\mathbf{h}) = \prod_{i=1}^{2^m} f(G_i|\mathbf{h}).$$

This can be done by the EM algorithm.

EM: toy example in R

```
> library(haplo.stats)
> snp1 <- c("AT","AT","AA","TT","AA")
> snp2 <- c("AA","GG","AG","GG","AA")
> snp3 <- c("CT","CT","CC","CT","TT")
> Geno <- cbind(substr(snp1,1,1),substr(snp1,2,2),
+               substr(snp2,1,1),substr(snp2,2,2),
+               substr(snp3,1,1),substr(snp3,2,2))
> snpnames <- c("snp1","snp2","snp3")
> HaploEM <- haplo.em(Geno,locus.label=snpnames,control=haplo.em.control(min.posterior=1e-4))
> HaploEM
=============================================================================================
                                          Haplotypes
=============================================================================================
  snp1 snp2 snp3 hap.freq
1    A    A    C      0.1
2    A    A    T      0.3
3    A    G    C      0.2
4    T    A    C      0.1
5    T    G    C      0.1
6    T    G    T      0.2
=============================================================================================
                                            Details
=============================================================================================
lnlike =  -14.18484
lr stat for no LD =  1.51763 , df =  2 , p-val =  0.46822
```

EM: empirical example of gene ACTN3 in R

```
>library(haplo.stats)
>fms <- read.delim(file="c:/data/FMS_data.txt",header=TRUE,sep="\t")
>attach(fms)
>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")
>Geno.C <- Geno[Race=="Caucasian" & !is.na(Race),]
> HaploEM <- haplo.em(Geno.C,locus.label=snpnames,control=haplo.em.control(min.posterior=1e-4))
> HaploEM
================================================================================
                                  Haplotypes
================================================================================
   actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064 hap.freq
1            C              A               C             G  0.00261
2            C              A               T             A  0.00934
3            C              A               T             G  0.01354
4            C              G               C             A  0.47294
5            C              G               C             G  0.01059
6            T              A               C             A  0.00065
7            T              A               T             G  0.39891
8            T              G               C             A  0.08557
9            T              G               T             A  0.00065
10           T              G               T             G  0.00520
================================================================================
                                   Details
================================================================================
lnlike =  -1285.406
lr stat for no LD =  2780.769 , df =  5 , p-val =  0
```

Phase estimation in R

```
> Geno.AA <- Geno[Race=="African Am" & !is.na(Race),]
> nrow(Geno.AA)
[1] 44
> HaploEMAA <- haplo.em(Geno.AA,locus.label=snpnames,control=haplo.em.control(min.posterior=1e-4))
> HaploEMAA
================================================================================
                                  Haplotypes
================================================================================
  actn3_r577x actn3_rs540874 actn3_rs1815739 actn3_1671064 hap.freq
1           C              A               C             A  0.01106
2           C              A               C             G  0.08130
3           C              A               T             G  0.03764
4           C              G               C             A  0.57763
5           C              G               C             G  0.01123
6           T              A               C             A  0.00065
7           T              A               T             G  0.17167
8           T              G               C             A  0.10833
9           T              G               C             G  0.00048
================================================================================
                                   Details
================================================================================
lnlike =  -84.9793
lr stat for no LD =  119.6062 , df =  4 , p-val =  0
>
```

Phase estimation in R

```
> HaploEM$nreps[1:5]        # Nr. compatible haplotype pairs.
indx.subj
1 2 3 4 5
1 2 2 2 1
> HaploEM$indx.subj[1:8]
[1] 1 2 2 3 3 4 4 5
> HaploEM$hap1code[1:8]     # Possible haplotypic constitutions.
[1] 4 3 7 8 7 8 7 4
> HaploEM$hap2code[1:8]
[1] 4 8 4 3 4 3 4 4
> hprob <- HaploEM$hap.prob
> hprob
 [1] 0.0026138447 0.0093400121 0.0135382727 0.4729357032 0.0105890282
 [6] 0.0006518550 0.3989126969 0.0855667219 0.0006548104 0.0051970549
> p1 <- 2*prod(hprob[c(3,8)])
> p1
[1] 0.002316851
> p2 <- 2*prod(hprob[c(4,7)])
> p2
[1] 0.3773201
> p1n <- p1/(p1+p2) # Posterior probability of (3,8)
> p1n
[1] 0.006102807
> p2n <- p2/(p1+p2) # Posterior probability of (4,7)
> p2n
[1] 0.9938972
> HaploEM$post[1:8]
[1] 1.000000000 0.006102808 0.993897192 0.006102808 0.993897192 0.006102808 0.993897192 1.000000000
>
```

Software for haplotype estimation

Haplotype estimation is and has been a topic of intense research, and there are many programs for it:

- PHASE (Stephens et. al., 2001)
- fastPHASE (Scheet & Stephens, 2006)
- BEAGLE (Browning et. al., 2007)
- IMPUTE2 (Howie et. al., 2009)
- SHAPEIT (Delaneau et al., 2011, 2012)
- ...

Many of these programs also:

- Estimate missing values
- Can infer intervening SNPs that have not been typed.

References

- Foulkes, A.S. (2009) *Applied statistical genetics with R*. Chapter 5. Springer.
- Neale, B.M. (2008) *Statistical genetics: gene mapping through linkage and association*. Chapter 17. Taylor & Francis.

Computer exercise

- Install the R packages `haplo.stats`.
- Load the database http://www-eio.upc.es/ jan/data/bsg/CHBChr2-2000.rda
- Remove all SNPs with one or more missing observations.
- Select the first 5 SNPs of the database.
- Prepare the data for the `haplo.em` function
- Estimate the haplotype frequencies.
- How many haplotypes are possible theoretically?
- How many haplotypes are found by the algorithm? Which haplotypes are most common?
- Now select the first 25 SNPs of the database.
- Prepare the data for the `haplo.em` function
- Estimate the haplotype frequencies.
- How many haplotypes are possible theoretically?
- How many haplotypes are found by the algorithm? Which haplotypes are most common?
- Did you find the same number of haplotypes as with 5 SNPs? Why or why not?
- On which factors does the number of haplotypes depend?