Haplotyping unrelated individuals

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Introduction

If parental genotypes are available, it is fairly straightforward to infer the haplotypes transmitted the offspring.

In most association studies, the individuals are all unrelated. Haplotypes must be either:

- Measured directly (usually expensive and time consuming)
- Inferred statistically

The problem

For the simplest case of two diallelic markers, there are 9 observable unphased genotypes, but 10 possible phased genotypes (4 haplotypes).

OBSERVED (Unphased)	Phased
A/A, B/B	AB/AB
A/A, B/b	AB/Ab
A/A, b/b	Ab/Ab
A/a, B/B	AB/aB
A/a, B/b	AB/ab or Ab/aB
A/a, b/b	Ab/ab
a/a, B/B	aB/aB
a/a, B/b	aB/ab
a/a, b/b	ab/ab

Complete disequilibrium

		rs4820268	
rs855791	A/A	A/G	G/G
A/A	0	0	25
A/G	0	75	7
G/G	40	12	1

There are no AA/AA, AA/AG, and AG/GA genotypes, but there are **AG**/AG, **GA**/GA and GA/**GG**. Therefore, we can infer that there are probably only 3 haplotypes segregating:

AG, GA and GG.

Complete disequilibrium 2

The rs8557918*A allele probably arose as a mutation in a founder who was rs4820268*G on the mutated chromosome. The two SNPs are close together, so no recombination event has yet broken up the disequilibrium.

Haplotype	Frequency
AA	0.0000
AG	0.4037
GA	0.5367
GG	0.0596

A set of adjacent SNPs that are in complete disequilibrium like this are an **LD block**. That is, LD blocks are separated by sites of **ancestral recombinants**.

“Gene-counting” or Expectation-Maximization

If there is less than complete disequilibrium, we can iteratively estimate the proportions of the two phased genotypes for the double heterozygotes.

	S Blood Group		
MN group	S/S	S/s	s/s
M/M	91	147	85
M/N	32	78	75
N/N	5	17	7

We start with a trial value for the haplotype frequencies:

$$x1 = P(MS) = \frac{1}{4}; \quad x2 = P(Ms) = \frac{1}{4}; \quad x3 = P(mS) = \frac{1}{4}; \quad x4 = P(ms) = \frac{1}{4}$$

D	*x1*	*x2*	*x3*	*x4*	M/M S/S	M/M S/s	M/m S/S	M/m S/s	M/m s/s	m/m S/S	m/m S/s	m/m s/s
0	0.25	0.25	0.25	0.25	33.56	67.12	67.12	134.25	67.12	33.56	67.12	33.56

From our trial haplotype frequencies, we obtained some expected counts (**E** step). Now we work out the next set of trial haplotype frequencies, based on these numbers (**M** step). For example,

$$x1_1 = 2\times 67.125 + 134.25 + 134.25 + p * 268.5$$

where $p = \frac{O}{1+O}$; and $O = \frac{x1_0 \times x4_0}{x2_0 \times x3_0}$. In this case, $p = \frac{1}{2}$.

We then obtain:

D	*x1*	*x2*	*x3*	*x4*	M/M S/S	M/M S/s	M/m S/S	M/m S/s	M/m s/s	m/m S/S	m/m S/s	m/m s/s
0.01	0.37	0.4	0.09	0.14	74.49	36.5	160.52	93.33	13.23	86.48	58.19	9.79

We keep repeating the same procedure. With each iteration, we get closer and closer to the correct values for the haplotype frequencies. We stop when the change from iteration to iteration is small enough.

Ten EM iterations

D	*x1*	*x2*	*x3*	*x4*	M/M S/S	M/M S/s	M/m S/S	M/m S/s	M/m s/s	m/m S/S	m/m S/s	m/m s/s
0	0.25	0.25	0.25	0.25	33.56	67.12	67.12	134.25	67.12	33.56	67.12	33.56
0.01	0.37	0.4	0.09	0.14	74.49	36.5	160.52	93.33	13.23	86.48	58.19	9.79
0.02	0.38	0.4	0.09	0.14	76.79	34.74	160.66	93.49	12.93	84.04	59.79	10.63
0.02	0.38	0.39	0.08	0.14	77.76	33.99	160.7	93.6	12.78	83.03	60.43	11
0.02	0.38	0.39	0.08	0.14	78.16	33.67	160.71	93.65	12.72	82.62	60.69	11.15
0.02	0.38	0.39	0.08	0.14	78.32	33.54	160.72	93.68	12.69	82.45	60.8	11.21
0.02	0.38	0.39	0.08	0.14	78.39	33.49	160.72	93.68	12.68	82.38	60.84	11.23
0.02	0.38	0.39	0.08	0.14	78.42	33.47	160.72	93.69	12.67	82.35	60.86	11.24
0.02	0.38	0.39	0.08	0.14	78.43	33.46	160.72	93.69	12.67	82.34	60.87	11.25
0.02	0.38	0.39	0.08	0.14	78.43	33.46	160.72	93.69	12.67	82.34	60.87	11.25

Uncertainty of haplotype inference

In the two diallelic marker situation, in eight of the nine cells of the table of genotypes, we can unequivocally work out the haplotypes underlying the genotype for each individual. For the double heterozygotes, we can only give a probability.

Commonly in older papers, the most likely haplotype for each individual was just imputed, then the data analysed as if this was the true haplotype. Obviously in cases where there are two haplotypes for an observed genotype at say 40% and 60% probability, the 40% probability haplotype would never appear in the analysis. This can lead to bias in some cases.

Multiple imputation is one simple way around this problem.

Extension to large numbers of SNPs

We can extend the method to quite large numbers of SNPs by applying the method in a stepwise fashion. We first produce haplotype frequencies for a pair of SNPs. We then estimate the disequilibrium between these haplotypes (which we are simply treating as alleles at a new "supermarker") and the next SNP. The resulting haplotypes are then combined with another SNP and so on. This approach was first implemented, I believe, by David Clayton in his **SNPHAP** program.

A more elaborate variant on this stepwise approach is the Partition-Ligation EM algorithm (**PLEM**).

In related approaches, population genetic models are incorporated into the model, which hopefully can better pick long haplotypes. A coalescent model is used to predict most likely haplotypes for a particular unphased genotype based on related haplotypes in the sample. **Phase** (Stephens 2001) is the prototypical program of this type, and is used in the HapMap project. This is an MCMC algorithm.

Software

There are a very large number of programs now available.

2SNP	Beagle
FastPhase	Gerbil
haplo.stats (in R)	Haploview
Haplotyper	HINT
HIT	Phase
PLEM	Shape-IT
SNPHAP	

Association of haplotypes to traits

For a categorical trait, this is a straightforward extension. We just estimate haplotype frequencies within each level of the trait, and test for equality of these frequencies across the levels, via a chi-square. For sparse tables (low counts of genotypes), we can perform simulation-based (eg permutation) tests.

For a quantitative trait, or a categorical trait with continuous covariates, we can carry out a regression analysis where instead of exactly known genotypes, we have to average over the possible phased genotypes for each individual. We use the probabilities of the different genotypes for each person to weight the contribution of that genotype to the regression.

Individual	Trait value	Unphased genotype	Phased genotype	Case Weight
1	14	M/M S/S	MS/MS	1
2	10	M/m S/s	MS/ms	0.6
2	10	M/m S/s	Ms/mS	0.4
3	22	M/m S/S	MS/mS	1

SNP tagging and imputation

Once we have haplotype frequencies, we can:

- Choose a subset of **tagging SNPs** on the haplotype
- Predict (**impute**) the genotype at a SNP based on other SNPs on the haplotype

Tagging SNPs allow one to estimate haplotype frequencies without genotyping all the SNPs making up the haplotype. This means less genotyping cost for the same amount of association information. **Haploview** offers a nice interface to a tagging algorithm.

SNP imputation is the use of tagging SNP genotype to predict the genotype at the other SNPs on the haplotype. It is especially useful if you are trying to replicate an association reported for a SNP by other authors, and only have data from neighbouring SNPs.

Many groups are using imputation to increase the number of SNP association tests in their GWAS from 500K or 1M to the 4M HapMap SNPs.

SNP imputation

Both Gudbjartsson et al (2008) and Brown et al (2008) reported association between SNPs on chromosome 20 and risk of cutaneous melanoma, but the SNPs involved were 100 kbp apart. The deCODE assocation involved a haplotype rs4911414-rs1015362, while the strongest Australian association was to rs4911442. We can use the deCODE haplotype to impute the rs4911442 genotype fairly precisely (data from the ALS 555K GWAS):

		rs4911442				
rs4911414	rs1015362	A/A	A/G	G/G	Prediction	Accuracy
G/G	A/G	10	0	0	A/A	100%
	G/G	104	11	0	A/A	90%
G/T	A/A	4	0	0	A/A	100%
	A/G	86	9	1	A/A	90%
	G/G	3	19	2	A/G	79%
T/T	A/A	14	2	0	A/A	88%
	A/G	4	5	0	A/G	56%
	G/G	0	0	1	G/G	100%