BMC Bioinformatics

Methodology article

SNP haplotype tagging from DNA pools of two individuals

Josephine Hoh[1], Fumihiko Matsuda[2], Xu Peng[2], Daniela Markovic[1], Mark G Lathrop[2] and Jurg Ott*[1]

Address: [1]Laboratory of Statistical Genetics, Rockefeller University, New York, NY 10021, USA and [2]Centre National de Génotypage, 91057 Evry, France

Email: Josephine Hoh - jhoh@linkage.rockefeller.edu; Fumihiko Matsuda - fumi@cng.fr; Xu Peng - pengxu@cng.fr; Daniela Markovic - markovic@linkage.rockefeller.edu; Mark G Lathrop - mark@cng.fr; Jurg Ott* - ott@rockefeller.edu

* Corresponding author

Published: 22 April 2003

BMC Bioinformatics 2003, **4**:14

Received: 23 November 2002
Accepted: 22 April 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/14

Abstract

Background: DNA pooling is a technique to reduce genotyping effort while incurring only minor losses in accuracy of allele frequency estimates for single nucleotide polymorphism (SNP) markers.

Results: We present an algorithm for reconstructing haplotypes (alleles for multiple SNPs on same chromosome) from pools of two individual DNAs, in which Hardy-Weinberg equilibrium conditions or other assumptions are not required. The program outputs, in addition to inferred haplotypes, a minimal number of haplotype-tagging SNPs that are identified after an exhaustive search procedure.

Conclusion: Our method and algorithms lead to a significant reduction in genotyping effort, for example, in case-control disease association studies while maintaining the possibility of reconstructing haplotypes under very general conditions.

Background

While SNPs in many ways are highly useful genetic markers, it may only be the joint effect of multiple SNPs in a gene that provides much information about association to disease because, taken together, the SNPs represent a much more polymorphic system than each SNP by itself. The effects of these SNPs are perhaps best represented by their haplotypes. Ideally, a researcher would want to obtain genotypes on *all* SNPs in a gene but this effort tends to be expensive. Thus, several authors have proposed that pools of DNA from n individuals be genotyped, which reduces genotyping costs by a factor of n [1]. In case-control association studies, an extreme approach is to form one pool for case and one pool for control individuals [2,3]. Originally, pooling was introduced as a method for efficiently screening for a rare disease [4] and thus for identifying individuals. On the other hand, grouping was later used to protect confidentiality [5].

Several methodological investigations of pooling efficiency have focused on marker allele frequency estimation [6,7] and on testing for association between markers and disease based on pooled data [8]. Extensions to two loci involve estimation of allele frequencies and the disequilibrium parameter based on pooled data [9].

Clearly, DNA pools of large numbers of individuals will only allow investigating SNP *alleles* and will be unable to look at haplotypes. A middle ground will be to form pools of small numbers of individuals. Such pools, for example, of $n = 2$ or $n = 3$ individuals are expected to lead to savings in genotyping costs while still allowing to recognize haplotypes. Here we present algorithms for inferring

Table 1: Pool phenotypes. "Assign?" indicates whether genotypes can be assigned to individuals in the pool, and "Homogeneous?" indicates whether the two individuals have the same genotype.

Phenotype	XXXX	XXXY	XXYY	XYYY	YYYY
Assign?	yes	yes	no	yes	yes
Homogeneous?	yes	no	–	no	yes

haplotypes from pools of $n = 2$ individuals. Our methods are applicable to any (reasonable) number of SNP markers (not limited to pairs of SNPs) and do not rely on the presence of Hardy-Weinberg equilibrium (HWE) as many other haplotyping approaches do, that is, they are largely parameter-free and data-driven.

Algorithm

Pool phenotypes

Consider a gene with s SNPs and a number m of pools (of n = 2 individuals). Denote the two alleles at a SNP by X and Y. Then five different phenotypes may be distinguished as shown in Table 1 (in addition to these five phenotypes there may be additional, more ambiguous situations, for example, that both X and Y alleles are known to be present but in unknown numbers).

Of the five phenotypes, all but one (XXYY) allow an unambiguous assignment of SNP genotypes to the two individuals in a pool although, of course, the order of individuals is unknown. For example, phenotype XXXY clearly identifies one individual as XX and the other as XY. Among the four phenotypes with unambiguous assignment, one can distinguish "homogeneous" and "heterogeneous" pool phenotypes, where the former exhibit two identical and the latter two different SNP genotypes in the two individuals. For those SNPs that allow an unambiguous assignment of genotypes, genotype-phase will be known (i.e., which individual has which genotypes at all SNPs) if at most one SNP is heterogeneous (the term genotype-phase as we introduce it here is different from gametic phase, which refers to haplotypes).

Resolving ambiguities and missing information

The phenotype XXYY is ambiguous in the sense that the two individual phenotypes could be XX and YY, or XY each. Also, some phenotypes may be missing altogether. Before we can proceed with reconstructing haplotypes we need to "fill in such holes". We do this by imputing (partially) missing information and assume that the missing piece of information is missing at random, that is, it may be imputed based on the distribution of this type of information in the remainder of the data.

Missing phenotype

For a given SNP, a pool may be missing a phenotype completely. In this case we proceed in one of two ways. If among the remaining pools one phenotype is at least twice as frequent as the next frequent phenotype, then we impute the most frequent phenotype. Otherwise we impute one of the phenotypes at random.

Unknown number of X and Y alleles

A pool may be known to contain X and Y alleles for a given SNP but their numbers may be unknown. In other words, the phenotypes may be XYYY, XXYY, or XXXY. Here we count the respective numbers n_X and n_Y of X and Y alleles in the other pools and impute the genotype XYYY if $n_X < n_Y$, XXYY if $n_X = n_Y$, and XXXY if $n_X > n_Y$. Scoring of alleles in pooled DNA, for example, by mass spectrometry (fluorescence intensity) has a limited resolution so that occasionally the specific number of X or Y alleles is unclear.

Resolving genotype-phase

As mentioned above, XXYY is the only phenotype not revealing genotype-phase, that is, the individual genotypes cannot be recognized. For such a pool, we look at those other pools that do reveal genotype-phase and count the number of individual genotypes as n_{XX}, n_{XY}, and n_{YY}. We assume genotypes (XX, YY) if $n_{XX} + n_{XY} > n_{YY}$ and assume genotypes (XY, XY) otherwise. If $n_{XX} + n_{XY} = n_{YY}$ then we make a random assignment based on equal probability of the two situations.

Because of the random element involved in some of these imputations they may to lead to different results at different times. Thus, it may be optimal to repeat these procedures a number of times and suitably combine results. Another option would be to allow for the missing observations by maximum likelihood (in the EM algorithm outlined below) but this would generally require consideration of an unsuitably large number of genotype vectors.

Estimating genotype vector frequencies

At this point, each pool unequivocally exhibits the SNP genotypes of each individual in the pool. We see the following six genotype pairs: XX/XX, YY/YY, XX/XY, XY/YY,

XX/YY, and XY/XY. Among these genotype pairs, we distinguish two types: The two individuals in a pool have the same or different genotypes. We call the former *homogeneous* pools and the latter *heterogeneous* pools. The individual genotypes at different SNPs form a genotype vector that is analogous to a haplotype, which is a set of alleles at different loci. We recognize individual genotype vectors if among all SNPs at most one shows a heterogeneous phenotype.

Most of the time there will be multiple SNPs exhibiting heterogeneous phenotypes. Thus, we cannot generally count genotype vectors but must estimate them. We do this by a maximum likelihood method. By estimating frequencies of genotype vectors we do not need to make any assumptions on Hardy-Weinberg equilibrium. For s SNPs, there are a total of 3^s different genotype vectors. For example, s = 15 leads to 14,348,907 possible genotype vectors. In order to avoid having to enumerate all these, we focus on the subset compatible with the observed pools. For each pool, we generate a list of all genotype vectors that are compatible with the given pool. Initially, each of the different genotype vectors is assigned the same frequency, which is updated iteratively by an EM-type algorithm. At each iteration, we compute the sum of squared differences between current and previous genotype vector frequencies. Iteration stops when this sum of squares falls below 10^{-11}.

Haplotype reconstruction

A number m of pools represent $2m$ genotype vectors (two for each individual in a pool). So, with given estimated frequencies, g_i, for the i-th genotype vector, we prepare the following list: $[2mg_i]$ is the estimated number of individuals with the i-th genotype vector, where $[r]$ represents the number r rounded to its nearest integer. For each genotype vector, its estimated frequency is converted into an estimated number of individuals with this genotype vector. From this list of individuals and their SNP genotypes, we now reconstruct individual haplotypes by Clark's method [10], which, again, does not assume Hardy-Weinberg equilibrium. At this point, of course, haplotypes may be reconstructed and their frequencies estimated with alternative approaches [11–14] but most of these methods assume Hardy-Weinberg equilibrium. Although Clark's method does not necessarily lead to unique solutions, our main aim here was to continue analysis with a method not requiring the assumption of Hardy-Weinberg equilibrium. To evaluate the variability of the resulting haplotype frequency estimates it is recommended to carry out the whole procedure multiple times.

Implementation

As described above, we developed our approach in the following three steps: (1) For those pools not unequivocally exhibiting individual genotypes for a given SNP, we impute these genotypes based on known genotypes in the data. After this step we know the two genotypes at each SNP but not which person has which genotype. (2) For a given set of SNPs in a gene, the SNP genotypes of an individual form a *genotype vector*. We set up an EM algorithm to estimate the frequencies of all genotype vectors compatible with SNP genotypes and in this manner find the generally small number of genotype vectors with non-zero frequencies. (3) Based on the resulting SNP genotypes, Clark's algorithm [10] is applied to infer individual haplotypes and an exhaustive search of subsets of SNPs [15] is conducted to find the smallest number of such marker loci that uniquely identify ("tag") haplotypes (see *add2* sample data, below), that is, that represent haplotype-tagging SNPs (htSNPs). These algorithms have been implemented in a computer program package *pools2* that is freely available to researchers. It contains the programs *snp* (for steps 1 and 2 above) and *htSNP.py*, and also the raw data for the *add2* gene mentioned below. The package may be downloaded from the web site, http://linkage.rockefeller.edu/register.

Results and Discussion

Efficiency of the *snp* program depends on the informativeness of the pool phenotypes. For example, a total of 25 SNPs and indels had originally been genotyped in the *add2* gene (data file contained in the *pools2* package mentioned above). But to evaluate all 25 or even only the first 20 of them resulted in over 100,000 genotype vectors compatible with pool phenotypes, which presumably would have required enormous computing times. Thus, analysis of the first 15 SNPs by the *snp* program is shown in the next paragraph. It required 31 iterations and an execution time of 13.7 minutes on a Dell PC (1.4 GHz clock speed, 786 MB of RAM, Windows 2000).

Analysis of add2 gene

For the first 15 SNPs in the *add2* gene genotyped on 16 pools (two individuals each) of DNA we applied the procedure outlined above. Of the 6,252 genotype vectors compatible with the data (pool phenotypes), 12 have non-zero estimated frequencies. Haplotype reconstruction by the Clark method leads to 11 different haplotypes in the 2 × 16 = 32 individuals (Table 2).

Some of the 15 SNPs are identical: s12 = s13 = s14 and s3 = s4. Thus, there are a total of 12 unique SNPs. Each haplotype forms a pattern of *0*s and *1*s at these SNPs. If for a subset of SNPs this pattern is different for each haplotype then that set of SNPs collectively tag the haplotypes and represent a set of htSNPs. To find the minimum number of haplotype tagging SNPs we proceed as follows [15]. Let h denote the given number of haplotypes and k the number of SNPs that tag these haplotypes. For example,

Table 2: 15 SNPs and 11 haplotypes in the *add2* gene. The two alleles at each SNP are labeled 0 and 1.

Frequency		SNPs														
abs.	relative	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15
23	0.3594	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0.2188	0	0	1	1	1	0	1	0	1	0	1	1	1	1	1
9	0.1406	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
8	0.1250	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0.0625	0	0	0	0	1	0	1	0	1	0	1	1	1	1	1
1	0.0156	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0.0156	0	1	1	1	0	0	0	0	0	0	0	1	1	1	1
1	0.0156	1	0	1	1	1	0	1	1	1	0	1	0	0	0	1
1	0.0156	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
1	0.0156	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0.0156	1	0	1	1	1	0	1	1	0	1	0	0	0	0	0
64	1															

two SNPs can generate at most four different allele patterns. Therefore, two SNPs cannot possibly tag more than four haplotypes. Generally, we must have $2^k \geq h$ or, equivalently,

$$k \geq \log(h)/\log(2). \quad (1)$$

To be certain that we find the smallest set of htSNPs we carry out an exhaustive search of all subsets of SNPs as previously proposed [15]. We begin with subsets of k SNPs, where k is the smallest integer satisfying equation (1). For each subset, the number t of different allele patterns is evaluated. If $t = n$ we have found a set of htSNPs. We wrote a computer program, *htSNP.py*, to carry out this exhaustive search. For the 15 SNPs in Table 2, the program finds $k = 6$ as the smallest set of htSNPs and identifies 10 such sets. When we collapse s13 and s14 into s12, and s4 into s3, the following three sets of htSNPs remain: (s1, s3, s6, s7, s8, s15), (s1, s3, s6, s8, s9, s12), and (s1, s3, s6, s8, s9, s15). Thus, to identify the haplotypes in the *add2* gene, genotyping effort can be reduced from genotyping 15 SNPs to genotyping only 6 SNPs. Which of these sets of htSNPs is optimal depends on the population haplotype diversity [15]; in the given sample, each set of htSNPs identifies all haplotypes perfectly but this is not necessarily the case in a new sample.

If we disregard the six haplotypes occurring only once we are left with five haplotypes comprising a total frequency of 90.6% (Table 2). For these common haplotypes, the only polymorphic and unique SNPs are s1, s3, and s11. They must be haplotype tagging because we need at least 3 SNPs to tag 5 haplotypes. Thus, to tag common haplotypes in the *add2* gene it is sufficient to genotype those three SNPs.

Searching for htSNPs among all SNPs in a gene is somewhat related to looking for a cladistic structure among the haplotypes [16]. The most polymorphic SNPs are likely to be the oldest and tend to be included in the set of htSNPs. Thus, htSNPs tend to comprise those SNPs closest to an original disease mutation. A note of caution – while a possibly small number of htSNPs can identify all haplotypes, using only the htSNPs rather than all original SNPs in a gene does lead to reduced power in case-control disease association studies [17].

Uncertainties

Our procedures allow for some types of errors, for example, uncertain numbers of X and Y alleles in a pool, while assuming absence of other errors such as sample swaps [18]. The described method for imputing missing data may be suboptimal as it works with only one SNP at a time. We are currently working on improved approaches that take information from neighboring SNPs into account because SNPs in a gene tend to be highly correlated.

It should be pointed out that heterozygosity at multiple SNPs potentially introduces errors into the estimation of genotype vectors. Thus, some genotype vectors will be known with certainty and others only probabilistically. Also, the same pools causing genotypic ambiguity may also lead to haplotype ambiguity even given known genotypes. At this stage, we don't have any good answers to these observations except that increased sample size is expected to reduce such uncertainties.

Authors' contributions

JO and JH developed and implemented the algorithm and drafted the manuscript, MGL suggested the problem, FM and XP contributed single nucleotide polymorphism data, and DM carried out analyses.

Acknowledgements

Support through NIH grants R01-MH59492, R01-HG00008, and K25-HG00060 is gratefully acknowledged. We also thank Aventis Pharmaceuticals for supporting this work, and an anonymous referee for pointing out references [1] and [4,5].

References

1. Sham P, Bader JS, Craig I, O'Donovan M and Owen M **DNA pooling: A tool for large-scale association studies** *Nature Reviews Genetics* 2002, **3:**862-871
2. Arnheim N, Strange C and Erlich HH **Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: Studies of HLA class II loci** *Proc Natl Acad Sci* 1985, **82:**6970-6974
3. Plomin R, Hill L, Craig IW, McGuffin P, Purcell S, Sham P, Lubinski D, Thompson LA, Fisher PJ, Turic D and Owen MJ **A genome-wide scan of 1842 DNA markers for allelic associations with general cognitive ability: a five-stage design using DNA pooling and extreme selected groups** *Behav Genet* 2001, **31:**497-509
4. Dorfman R **The detection of defective members of large populations** *Ann Math Stat* 1943, **14:**436-440
5. Gastwirth JL and Hammick PA **Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood** *J Stat Planning Inference* 1989, **22:**15-27
6. Jawaid A, Bader JS, Purcell S, Cherny SS and Sham P **Optimal selection strategies for QTL mapping using pooled DNA samples** *Eur J Hum Genet* 2002, **10:**125-132
7. Shaw SH, Carrasquillo MH, Kashuk C, Puffenberger EG and Chakravarti A **Allele frequency distributions in pooled DNA samples: Applications to mapping complex disease genes** *Genome Res* 1998, **8:**111-123
8. Risch N and Teng J **The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling** *Genome Res* 1998, **8:**1273-1288
9. Pfeiffer RM, Rutter JL, Gail MH, Struewing J and Gastwirth JL **Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium** *Genet Epidemiol* 2002, **22:**94-102
10. Clark AG **Inference of haplotypes from PCR-amplified samples of diploid populations** *Mol Biol Evol* 1990, **7:**111-122
11. Chiano MN and Clayton DG **Fine genetic mapping using haplotype analysis and the missing data problem** *Ann Hum Genet* 1998, **62:**55-60
12. Zhao JH, Curtis D and Sham PC **Model-free analysis and permutation tests for allelic association** *Hum Hered* 2000, **50:**133-139
13. Stephens M, Smith NJ and Donnelly P **A new statistical method for haplotype reconstruction from population data** *Am J Hum Genet* 2001, **68:**978-989
14. Lin S, Cutler DJ, Zwick ME and Chakravarti A **Haplotype inference in random population samples** *Am J Hum Genet* 2002, **71:**1129-1137
15. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA and Dudbridge F **Haplotype tagging for the identification of common disease genes** *Nat Genet* 2001, **29:**233-237
16. Templeton AR, Weiss KM, Nickerson DA, Boerwinkle E and Sing CF **Cladistic structure within the human Lipoprotein lipase gene and its implications for phenotypic association studies** *Genetics* 2000, **156:**1259-1275
17. Zhang K, Calabrese P, Nordborg M and Sun F **Haplotype block structure and its applications to association studies: Power and study designs** *Am J Hum Genet* 2002,
18. Kirk KM and Cardon LR **The impact of genotyping error on haplotype reconstruction and frequency estimation** *Eur J Hum Genet* 2002, **10:**616-622