SNP cherry picker: maximizing the chance of finding an association with a disease SNP

Mark Harris*, Jeremy M. R. Martin, John F. Peden and Christopher J. Rawlings

Oxagen Ltd, 91 Milton Park, Abingdon, Oxon OX14 4RY, UK

Received on Feb 24, 2003; revised on April 8, 2003; accepted on April 26, 2003

ABSTRACT

Summary: The high cost of genotyping single nucleotide polymorphisms (SNPs) generally prohibits the systematic mapping of entire genetic linkage regions in order to find the polymorphisms associated with increased risk of disease. In practice, SNPs are selected at approximately equal spacing across the linkage region to try to locate a SNP lying in the haplotype block of the disease SNP. The size of the haplotype block may not be known, however, and SNPs taken from public domain sources may not in fact be polymorphic. Our program will choose a subset of the SNPs in a linkage region so as to maximize the expected proportion of the sequence that lies within a given distance of a real SNP.

Availability: The software is available, free of charge, for academic use on request from the authors.

Contact: SD@oxagen.co.uk

Supplementary Information: www.oxagen.co.uk

A typical approach to finding the cause of a genetic disease involves using linkage studies on genetic markers to determine regions associated with the disease (e.g. Rioux et al., 2001; Eerewegh et al., 2002). These linkage regions are then mined for SNPs and a subset of these SNPs are genotyped in the sample population. The aim is to locate the haplotype block of the disease SNP (Johnson et al., 2001). These blocks are regions of DNA with very little diversity interspersed with recombination sites. The SNPs in a haplotype block are usually all in very high Linkage Disequilibrium (Daly et al., 2001). The SNPs in this region are more likely to have an association with the disease and can be investigated further as causes of missense or promoter mutations.

The scale of this approach can be overwhelming; large linkage regions can contain thousands of possible SNPs. In a genotyping project with limited resources, SNPs need to be chosen carefully from across the whole region to improve the likelihood of finding the disease SNP through association. Since the haplotype map of the human genome is patchy, we are unsure as to the size of haplotype blocks of each SNP, and therefore our approach is to maximize the amount of the sequence lying within a given distance of a real SNP. This problem is made more difficult because data in the public domain databases are of variable quality and the coverage will not be consistent. Furthermore, many SNPs from these databases are unlikely to be real mutations (Marth et al., 2001; Small et al., 2002), and may be attributable to human or sequencing errors. We have assigned probabilities to each putative SNP in the linkage region of interest representing how likely the SNP is to be genuine. These probabilities were based on the source and/or detection method for each SNP as reported by dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). In cases where we had already genotyped SNPs from a given source and/or method, we used the fraction of SNPs we had empirically found to be polymorphic in our sample populations. In other cases, we assigned a conservative estimate of $p = 0.40$ (Marth et al., 2001).

We have modelled the problem as follows: Formal statement of the SNP coverage problem. We consider a DNA sequence $SEQ$ containing some number $n$ of putative SNPs: $S_1, S_2, \ldots, S_n$, each in some particular position along the sequence. Each SNP $S_j$ has a certain probability $P_j$ of being ‘real’. We assume these probabilities are independent.

We say that a base position along the sequence is ‘covered’ if it lies within a fixed range $R$ of the nearest real SNP.

The ‘coverage’ of a sequence by a set of real SNPs is then defined as the proportion of its bases that are covered by the SNPs.

Our aim is to choose a subset of exactly $m$ putative SNPs, $T_1, T_2, \ldots, T_m$, so as to provide the best expected coverage of $SEQ$.

Consider base position $i$ along sequence $SEQ$. The probability that position $i$ is covered by a given set of putative SNPs is calculated in the following way:

Suppose that there are three putative SNPs: $S_1, S_2, S_3$, within distance $R$ of $i$. The probability that none of these SNPs is real is

$$1 - P_1 - P_2 - P_3$$

Hence the probability that position $i$ is covered is

$$1 - (1 - P_1)(1 - P_2)(1 - P_3)$$

*To whom correspondence should be addressed.
Note that this is a step function, which changes value only when a particular SNP goes in or out of range \( R \) from base \( i \).

We can calculate the expected coverage of sequence \( SEQ \) for any given subset of putative SNPs by integrating the coverage probability along the sequence.

This problem can be summarized in the following objective function:

Let \( I \) be the set of bases in the sequence and \( J \) the set of SNPs under consideration. Let \( P_{ij} \) be the probability that base \( i \) is covered by SNP \( j \). We define \( x_j = 1 \) if SNP \( j \) is included in the set of chosen SNPs and \( x_j = 0 \) otherwise.

Then the maximum expected coverage of \( k \) SNPs is given by:

\[
\text{Max} \sum_{i \in I} \left[ 1 - \prod_{j \in J} (1 - P_{ij}x_j) \right] \quad (1)
\]

s.t. \[
\sum_{j \in J} x_j = k \quad (2)
\]

\[
x_j \in \{0, 1\} \quad j \in J \quad (3)
\]

A mathematically equivalent problem has been described in the field of Environmental Studies (Camm et al., 2002). This was proved to be an NP-hard problem, and hence there is little hope of finding an efficient algorithm to solve this problem. A proof of this fact is given in the on-line supplementary paper.

The SNP cherry picker algorithm. We desire to find the subset of \( m \) SNPs from \( n \), which gives the greatest expected coverage of the sequence \( SEQ \). There are \( \binom{n}{m} \) such subsets, which is likely to be an infeasibly large number to explore exhaustively. Our approach is to use a greedy algorithm (Cormen et al., 1990) to find a solution that is a close approximation to this true optimal one.

The algorithm starts with the complete set of SNPs. One SNP is removed from the selected set at a time so as to maximize the expected coverage of the remaining set of SNPs. This approach has \( o(n^2) \) complexity.

Figure 1 shows a comparison between three methods for selecting SNPs—the greedy algorithm, a random approach and the naive approach of choosing SNPs at regular intervals purely based on position. The graph plots the expected coverage against the number of SNPs chosen. The initial set contained 30,000 SNPs from a chromosomal region approximately 22 Mb long. The coverage range \( R \) was set to 2.5 kb, which was well within current estimates of the size of haplotype blocks (Daly et al., 2001).

Firstly, we see that the maximum expected coverage is only 86%. This is because (i) there are significant stretches of the sequence where no SNP data exists and (ii) unsubstantiated public domain SNPs were only given low probabilities. We can see that choosing only a set of 5500 SNPs out of 30,000 using our greedy algorithm gives an expected coverage of about 80% of the sequence. The set chosen by taking SNPs based purely on their position would only give a coverage of 70%, and to produce the same expected coverage would require choosing 8600 SNPs. Although at first sight, this seems a relatively modest improvement, the advantage becomes clear when the costs of a full SNP genotyping experiment are considered. Such a study might genotype 1000 cases and 1000 control samples. Given that the cost of genotyping a single SNP now costs approximately $0.20, this represents a saving of approximately $1.24M.

This approach assumes that the probabilities assigned to the SNPs are meaningful and correct. We used very conservative estimates for our success probabilities in cases where confidence was low because only a small number of SNPs were previously genotyped for a given method. The probabilities could also be reinterpreted to address other aspects of SNP selection. One such application might be to represent the usefulness of a particular assay if different genotyping technologies are available, for example.

As an alternative approach, Camm et al. (2002) used a linear programming algorithm to find an approximate solution to the analogous problem of optimizing the species diversity when selecting possible sites for nature reserves. This would be an interesting algorithm to study for future development.

Implementation. Our program was developed entirely using Java. It takes a tab-delimited text file in the following format: ‘<Location> <SNP Name> <Probability> <Allele Frequency>’. Location is the position of the SNP in genomic coordinates. Allele Frequency is the lower of the two frequencies for the alleles of the SNP. Coverage range \( R \)
SNP cherry picker algorithm

can be set at runtime. The output is printed in two columns, the name of the SNP removed and the expected coverage at that point. Details of the Java classes are given in the on-line supplementary paper.

REFERENCES


