Maximizing power in association studies

Eran Halperin & Dietrich A Stephan

Only a subset of genetic variants can be examined in genome-wide surveys for genetic risk factors. How can a fixed set of markers account for the entire genome by acting as proxies for neighboring associations?

The etiology of many complex diseases is attributed to a combination of genetic and environmental risk factors. Knowledge of these influences yields insight into disease mechan isms and can thus ultimately enable better preventive, diagnostic and therapeutic strategies. The most common genetic variants in the human genome are single nucleotide polymorphisms (SNPs)—point mutations with multiple possible alleles at a locus across the population. Genome-wide association (GWA) studies examine the set of cases and controls at many polymorphic sites and often identify one or several physical location(s) in the genome where genetic variation contributes to disease susceptibility[1].

Although conceptually straightforward, the statistical and computational aspects of GWA studies are considerable. They encompass the design of a well-powered study, controlling for confounding risk factors (e.g., population structure or exposure to environmental risks), accurate genotyping, correcting for multiple hypothesis testing and defining interactions between different SNPs. Such statistical measures are necessary whether we use current high-density SNP genotyping approaches or complete whole-genome sequencing in the future. We discuss here the foundation that allows us to capture information about regions of the genome that are currently not genotyped using standard high-throughput technologies. Understanding these computational approaches is key to maximizing identification of disease-associated DNA variants.

Eran Halperin is at the International Computer Science Institute, Berkeley, CA 94704, USA, and the Computer Science and Biotechnology Departments, Tel-Aviv University, Tel-Aviv, 69978, Israel. Dietrich A. Stephan is in the Division of Genomics Research, Navigenics, 1001 E. Hillsdale Blvd., Foster City, CA 94404, USA. e-mail: dietrich@navigenics.com

Indirect association and linkage disequilibrium

Recent technological advances allow us to rapidly genotype $>10^6$ SNPs in an individual, accounting for 10% of the estimated number of common SNPs (>1% minor allele frequency) across the population[2]. As a result, true associations might be missed if the causal SNP is not genotyped or if the causal variant is an unknown variant. Computational methods have been developed to account for some of the unobserved variants[3–7]. The rationale for these methods is based on the observation that SNPs in close proximity to one another in the genome tend be correlated, or in linkage disequilibrium.

There are a few metrics that measure the linkage disequilibrium between a pair of SNPs. The linkage disequilibrium parameter D measures the linkage disequilibrium between a pair of SNPs s_1 and s_2. D is defined as $D = P_{12} - p_1p_2$, where P_{12} is the frequency of chromosomes with the minor allele present in both SNPs, and p_i is the frequency of the minor allele frequency at SNP s_i. Intuitively, D measures the deviation of the joint distribution from the case where the SNPs are inherited independently. It is largely determined by the recombination rate between the two SNPs. If ρ is the probability of a recombination in a single meiosis in the region spanned by these SNPs, the linkage-disequilibrium parameter should change to $D_n = (1-\rho)D_{n-1}$ in subsequent generations. A more commonly used metric is $D' = D/D_{\max}$, where $D_{\max}$ is the maximal possible value of D for the given allele frequencies p_1 and p_2. As this metric does not directly depend on the allele frequencies, we can compare 'apples to apples' when contrasting linkage disequilibrium between different pairs of SNPs.

In association studies, linkage disequilibrium between SNPs can be used to replace a direct association test with an indirect one (tagSNP). As current technology does not allow us to genotype all known SNPs, we pick a set of tagSNPs such that the ungenotyped SNPs (or hidden SNPs) are in linkage disequilibrium with the tag SNPs. Thus, if the causal SNP is a hidden SNP, we expect to find a correlation between the phenotype and the tag SNPs due to correlation between the two SNPs. To do so, we first have to decide on a criterion for when one SNP 'captures' another. Although D' is a possible candidate, the relation between D' and the power to detect association is not clear. Alternatively, one can simply measure the correlation coefficient r between the two SNPs. The correlation coefficient is a measure, which ranges from –1 to 1, of how well two variables predict each other; formally, it is defined as

$$r = \frac{D}{\sqrt{p_1(1-p_1)\,p_2(1-p_2)}}$$

Often, the square of the correlation coefficient is used; whereas $r^2 = 1$ indicates that the two SNPs are perfectly correlated, $r^2 = 0$ implies that the two SNPs segregate independently throughout the population. The correlation coefficient is often chosen as the criterion for selecting tagSNPs, as there is a direct relation between r^2 and the power to detect association. If the true causal SNP is s_1, then the power to detect association at s_2 by genotyping N individuals is approximately the power attained by genotyping r^2N individuals at s_1 (ref. 8).

Based on this observation, an ideal set of tagSNPs will be a minimal set of SNPs with a high correlation coefficient between every hidden SNP and its corresponding tagSNP. The definition of 'high' may be somewhat subjective, and it generally depends on the resources available (that is, the total number of SNPs that will be genotyped). As the power to detect association in SNPs depends on their allele frequency, it is advised to use a more stringent threshold for such SNPs. In practice, however,

Table 1 Haplotypes improve the prediction of hidden SNPs

SNP1	SNP2	SNP3	Frequency
A	A	C	23%
A	A	T	1%
G	A	C	40%
G	G	C	2%
G	G	T	34%

SNPs 1 and 2 alone have poor power to predict the genotype at SNP 3, even when the phased haplotype is known. But together, using a multimarker tagSNP approach, SNPs 1 and 2 predict SNP 3 with 97% accuracy.

association studies are normally designed with a fixed threshold in mind for all SNPs; a common choice is a threshold of $r^2 > 0.8$.

Unfortunately, defining the best set of tagSNPs is computationally intractable in its full general form. In practice, an iterative greedy algorithm works well[3]. This algorithm analyzes a reference data set, such as the data provided by the International HapMap Project, in which 270 individuals from four different populations were genotyped at 3.1 million SNPs across the genome[9]. The algorithm finds a set of tagSNPs that 'covers' all other SNPs, where SNP s_1 covers SNP s_2 if the r^2 between them is larger than a threshold specified by the user. The algorithm works in iterations; initially, all the genotypes of the SNPs in the reference data set (in this case 3.1 million) are considered 'uncovered'. An iteration involves finding a tagSNP that covers the maximum number of uncovered SNPs. The tagSNP, as well as the SNPs that it covers, is considered covered from that point further. The algorithm ends when all possible SNPs are covered. This method is effective and widely used to define linkage-disequilibrium structure and tagSNPs in the genome. The scaffold can then be superimposed on the available high-density genotyping platforms, and the subset of hidden SNPs that the platform captures can be identified.

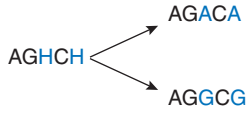

Figure 1 TagSNPs and haplotype information can enhance the ability to identify disease-related loci using linkage disequilibrium. On the left side is a genotype of an individual, where A,G,C,T correspond to homozygous genotypes and 'H' denotes a heterozygous genotype. On the right side are the two haplotypes of the same individual. Another possible pair of haplotypes that explain the genotype is AGACG, AGGCA. Phasing methods use the population information to infer which of the possible haplotypes is correct.

Multimarker methods and haplotypes

We have discussed the possibility of having one tagSNP that covers a hidden SNP. Often, multiple tagSNPs serve as a better proxy for a hidden SNP than any single tagSNP. In **Table 1**, SNPs 1 and 2 cannot serve as a proxy to SNP 3, but together, they correlate almost perfectly to SNP 3 (that is, when SNPs 1 and 2 carry the A allele, then SNP 3 most likely carries the C allele). We can thus predict SNP 3 with a 3% error rate by considering only tagSNPs 1 and 2—a considerably better outcome than when using SNP 1 or 2 alone. Note, however, that we are using the haplotype information and not genotypes. Unlike genotypes, which represent the allelic information on both chromosomes, haplotypes represent the information on only one of the chromosomes (**Fig. 1**).

Current whole-genome platforms can genotype a fixed set of SNPs that cannot be customized per experiment. To take advantage of haplotypes within these constraints, de Bakker *et al.* suggested that for every hidden SNP *s*, one can exhaustively search the HapMap data set for a proxy haplotype for which the square of the correlation coefficient with SNP *s* is higher than a given threshold[4]. Deriving a haplotype proxy is not a computationally trivial task, as the number of potential haplotypes is enormous. In principle, every set of SNPs (not necessarily consecutive) may potentially span a haplotype proxy. Exhaustively searching across all possible sets of SNPs is infeasible; however, to allow for a manageable running time, the algorithm considers only short haplotypes (2–3 SNPs) and only SNPs in close proximity to the hidden SNP. As SNPs that are physically far from the hidden SNP are unlikely to correlate well with it owing to increased probability for recombination between the sites, these restrictions do not cause substantial loss of information. Once the proxy is found, the haplotype can be tested for association with the disease by performing a standard χ^2 test. de Bakker *et al.* have shown that the use of haplotypes is beneficial and consequently increases the power to detect an association[4]. Intuitively, this is because the number of haplotypes in any given region is smaller than the number of genotypes (**Table 2**), resulting in a larger sample size that is used to estimate any given haplotype. More importantly, the haplotypes represent the ancestral genetic structure that is shaped by evolutionary forces such as recombination rates and mutations, and these are implicitly taken into account when haplotypes are analyzed, as opposed to genotypes.

The above discussion deals with the case where the set of genotyped SNPs is not necessarily fixed. However, in practice, high-throughput genotyping platforms are designed so that there is no flexibility in the tag SNP selection.

Table 2 Genotype prediction power

SNP1	SNP2	SNP3	Frequency
A	A	C	5.3%
A	A	H	0.5%
H	A	C	18.4%
H	H	C	0.9%
H	H	H	15.7%
H	A	H	0.8%
H	H	T	0.7%
G	A	C	16%
G	H	C	1.6%
G	H	H	27.2%
G	G	H	1.3%
G	G	T	11.6%

When using genotype information, the same SNPs 1 and 2 have less power to predict the genotype at SNP 3, as the third SNP remains ambiguous even when the full genotype information is given at SNPs 1 and 2. 'H' denotes a heterozygote for that SNP.

Recently, different approaches have been proposed to choose a set of haplotype-based statistical tests that will be performed on the data given a fixed set of tagSNPs. One generalization of the haplotype-based test assigns a weight w_i to each haplotype h_i, and the resulting proxy for a nearby SNP is given by $\sum_i w_i h_i$ (ref. 5). An optimal choice of the weights guarantees improved power compared to the single-SNP or single-haplotype tests as these tests correspond to specific choices of the weights. It turns out that such an optimal set of weights corresponds to the probabilistic 'imputation' of a hidden SNP using the observed SNPs; in other words, we can use the haplotype structure of a reference population such as the HapMap[9] to learn the conditional distribution of a hidden SNP based on the haplotype distribution in the tagSNPs. Currently there is a major effort to improve the methods for imputation of hidden SNPs, as these methods promise to improve the power of association studies and to reach SNPs that have not been genotyped in the study. We will discuss these methods and their applications in genome-wide association studies in a future paper.

ACKNOWLEDGMENTS

E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel Aviv Univerity.

1. The Welcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
2. Kruglyak, L. & Nickerson, D.A. *Nat. Genet.* **27**, 234–236 (2001).
3. Carlson, C.S. *et al. Am. J. Hum. Genet.* **74**, 106–120 (2004).
4. de Bakker, P.I. *et al. Nat. Genet.* **37**, 1217–1223 (2005).
5. Zaitlen, N., Kang, H.M., Eskin, E. & Halperin, E. *Am. J. Hum. Genet.* **80**, 683–691 (2007).
6. Marchini, J. *et al. Nat. Genet.* **39**, 906–913 (2007).
7. Minichiello, M.J. & Durbin, R. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
8. Pritchard, J.K. & Przeworski, M. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
9. The International HapMap Consortium. *Nature* **449**, 851–861 (2007).