Population-based gene discovery in the post-genomic era

Naomi B. Zak, Sagiv Shifman, Anne Shalom and Ariel Darvasi

The complex genetic nature of many common diseases makes the identification of the genes that predispose to these ailments a difficult task. Consequently, many factors have to be considered in choosing the optimal approach to be taken in gene discovery of susceptibility genes. The elements to be considered include the applicability of a family-based linkage paradigm versus a population-based association design and the effects of linkage disequilibrium (LD) and genotypic relative risk (GRR). In this review we discuss these various points and describe the impact on LD and GRR of studying an isolated (also termed 'founder' or 'homogeneous') population, such as Ashkenazi Jews, as compared to an outbred population, such as Caucasians.

***Naomi B. Zak**
and Ariel Darvasi
IDgene Pharmaceuticals
PO Box 34478
Jerusalem 91344
Israel
*tel: +972 2 659 5600
fax: +972 2 659 5601
e-mail: naomiz@idgene.com
Sagiv Shifman
Anne Shalom
and Ariel Darvasi
The Life Sciences Institute
The Hebrew University of Jerusalem
Jerusalem 91904
Israel

▼ Although monogenic diseases often exhibit severe clinical phenotypes, the major burden of genetic ill health is caused by the more prevalent polygenic disorders, such as diabetes, hypertension, heart disease, cancer and others. These conditions affect millions of individuals and their management consumes vast amounts of healthcare resources. Yet, even after the Human Genome Project has completed a rough draft of the entire genome, the causative genetic variants of these ailments are still not known, and thus concerted efforts on the part of individuals working in both academia and biotechnology are required before pharmaceutical companies can make use of this knowledge to develop better therapies.

Why is this so? The explanation for these difficulties lies in the fact that cardiovascular disease, diabetes, mental illness, various cancer syndromes, Alzheimer's disease and a long list of other common ailments are all complex genetic diseases. That is to say, despite the overall heritability of the phenotype, the etiology of these diseases involves many different, potentially interacting, genetic and environmental risk factors; specific variants in any single gene are only contributory to the overall predisposition of a person to eventually suffer from that disease. This reduces the genotype–phenotype correlation on which gene discovery is based.

Nonetheless, what strategies are available for attempting to identify the genes that predispose to complex illnesses? Because no probable disease model can be assumed, model-independent methods of analysis are required. The model-independent linkage strategy that is most frequently employed is the sib-pair method, in which identity by descent (IBD) relationships are used in examining allele sharing among sibs (see Ref. 1 for theory of IBD and complex diseases). An alternative approach is that of allelic association[2]. Allelic association refers to a significantly increased or decreased occurrence of a marker allele in correlation with a disease trait. It can be because of either a true biological action of the examined polymorphism or because this polymorphism is in linkage disequilibrium (LD) with a nearby susceptibility gene. Because this approach is prone to false positives generated by various sources, as discussed below, the results obtained must be considered with care[2].

Association studies require highly dense sets of polymorphic markers that can be rapidly typed on large numbers of samples. The paucity of markers and lack of appropriate genotyping technology have been the major limitations for applying this strategy until now. However, a solution to this problem has recently presented itself when, in the course of the Human Genome Project, more than 1.68 million Single Nucleotide Polymorphisms (SNPs) were identified and made available to the scientific community (see http://www.ncbi.nlm.nih.gov/SNP/ and http://snp.cshl.org/). This number is expected to increase further.

1359-6446/01/$ – see front matter PII: S1359-6446(01)02007-4

SNPs represent the most common form of sequence variation. These single-base substitutions dispersed along the genome cause any two random human genomes to be 0.1% different from one another, and account for most of the heritable variation among individuals, including susceptibility to disease. SNPs are stable genetic markers, with a relatively low mutation rate. Being in most cases bi-allelic, they are also amenable to automatic genotyping.

In order for the tremendous potential of SNPs in association studies (and later in diagnostic tests) to be realized, technologies permitting rapid, accurate and inexpensive genotyping of multiple individuals with numerous markers must become available. Indeed, as discussed in several recent reviews[3,4], such high-throughput technologies are being rapidly developed, and their availability significantly increases the feasibility of population-wide allelic association studies.

Successes and limits of linkage analysis

Linkage analysis is an approach that has been widely employed in the past to identify disease genes. Linkage studies employ family samples to compare the segregation patterns of mapped genetic markers with that of the disease state. If any marker variant tends to be inherited together with the disease in question, this implies that the marker and the gene responsible for the phenotypic trait reside in physical proximity to each other. As the genetic markers have been mapped, they point to the approximate chromosomal location of the nearby disease gene.

Linkage analysis has been very successful in mapping several hundred Mendelian disease genes. Mutations in these genes lead, by themselves, to a disease phenotype. There are also several examples of common genes whose roles in complex, polygenic, diseases have been detected by these approaches. One is human leukocyte antigen (HLA). This was found to play a role in type I diabetes when it was observed that affected sib pairs (ASPs) share the allele in 73% of cases (instead of the expected 50%) (Ref. 5). Another is apolipoprotein E (ApoE), which is a major determinant in late-onset Alzheimer's disease[6].

These examples of genes whose roles in complex diseases have been unravelled by linkage analysis involve alleles with a very large relative effect. These are genes for which, despite the polygenic nature of the disease, there is a large difference in the average phenotype of individuals homozygous for each allele form, thus making the inheritance pattern more similar to Mendelian patterns. However, linkage analysis has not been successful in identifying genes of the kind that are likely to account for most of the genetic effects in complex diseases. These genes have smaller effects on the phenotype, and as a result a greater extent of allele sharing, or overlap, may be expected among individuals with different phenotypes. The reason that effects of small genes cannot be recognized by linkage analysis is that, in most family based studies, a restricted pedigree size severely limits the statistical power to detect the underlying genes[7].

Association analysis – advantages of the case-control paradigm

Population-wide direct-association analysis offers significantly higher statistical power than linkage analysis for the detection of genes with modest phenotypic effects[8]. The approach often used in such studies is the case-control design, which compares unrelated, affected individuals and unrelated, unaffected controls. In association analyses, polymorphic markers are tested for differences in allele frequencies between cases and controls. Enrichment of one allele form in diseased individuals is taken as evidence of association.

There are several advantages in employing the association study strategy for finding the genetic changes that underlie disease etiology. First, because of its greater statistical power, these studies provide a feasible paradigm to decipher the roles of genes with moderate effects on disease susceptibility, like the genes involved in polygenic diseases. In addition, the greater ease of matching cases to unrelated controls without the necessity of finding and enrolling additional family members to a study makes the case-control design a more economically attractive approach. Finally, population-wide association analyses have great potential for accurate mapping and identification of the actual gene as compared to family-based studies, which can detect only broad chromosomal regions.

Regardless of these advantages, several issues must still be addressed before the utility of this strategy is ascertained. The first of these regards LD. As described below, association analyses rely on LD, or a nonrandom correlation, between the polymorphic markers that are examined and unknown nearby disease-predisposing variants. Thus LD structure across the genome will determine the required marker density. What is the average distance over which LD operates? How does this extent range across the genome? How does it vary between populations?

A second issue is that of genotype relative risk (GRR). Undoubtedly, the association paradigm is better able to deal with genes of moderate input than are linkage studies. But what are the smallest effects they can be expected to reveal? What sample sizes will be required to detect these effects?

Finally, as mentioned above, a drawback of the association approach is the potential for false positives. One

source of spurious associations is inadequate ethnic matching of cases and controls and ethnic variation in allele frequencies. A second source might be inadequate correction for multiple testing. Assuming that the appropriate statistical measures are taken to correct for multiple testing, what steps can be taken to avoid false positives caused by ethnic stratification? One approach is to use unlinked genetic markers and statistical methods to infer details of population structure before testing for gene-disease associations. This approach has been described by Pritchard *et al.*[9] Another approach, as well as the issues regarding LD and GRR that were raised above, will be discussed.

Linkage disequilibrium and association analyses

In association analysis a high density of markers is employed and it is assumed that some of them will be in LD with the disease susceptibility gene or, less frequently, that the actual functional polymorphism will be tested. The extent of LD is largely determined by the distance of the polymorphic marker from the susceptibility gene. The shorter this distance, the fewer recombination events will occur between the two, and the more instances in which the susceptibility allele will not have been shuffled away from the marker form originally present on the chromosome where it first appeared.

Some theoretical and experimental observations have suggested that blocks of significant LD extend only a few kb around common polymorphisms, such as those expected to be involved in common diseases[10]. If this were really the case, it would mandate a very high density of polymorphic markers and make association studies impractical. However, in the first systematic, large-scale study of its kind, LD was found to typically extend 60 kb from common alleles in 19 randomly selected genomic regions[11].

It would be wrong, however, to oversimplify the issues. Although LD blocks are generally large, a high level of variation in LD was observed between the genomic regions studied. Not unexpectedly, variation was found to be correlated with the estimated local recombination rate[11]. These observations suggest that the optimal marker density for association studies will vary, depending on the genomic region under examination, and that a prerequisite to the planning of association studies is more knowledge about LD levels across the entire genome.

In addition, it was shown that not all populations display equivalent average values of LD. Specifically, a markedly shorter range of LD, extending less than 5 kb, was demonstrated in Nigerians. The differences between populations are thought to result from population history, probably a bottleneck effect that occurred among the ancestors of the North Europeans in contrast to the African population, which underwent simple expansion[11]. Thus, as will be discussed further below, the designers of association studies must also determine their population of choice. One interesting suggestion that has been raised is that genome-wide LD mapping be performed in two steps[11]. The first step would be with populations exhibiting maximal levels of LD in order to facilitate the initial identification of associated regions. In a second step, populations with much smaller blocks of LD would be utilized for high-resolution localization of the actual gene.

GRR and association analyses

The GRR is the ratio of the risk of developing the disease among individuals with the susceptibility genotype compared with the risk for individuals with an alternative genotype. Although they are not conventionally considered in this manner, the GRR for genes causing disorders inherited in a Mendelian manner may approach infinity. In stark contrast, the anticipated GRR values for genes that influence predisposition to complex diseases are in the single digit range. In the case of such weak gene effects, linkage analysis requires prohibitively large samples to provide evidence for the location of a disease gene.

To take, for example, the case of a yet-to-be-discovered disease-predisposing gene with a GRR of three and a standard set of assumptions, the required sample size under the sib-pair paradigm to detect involvement of this gene is calculated to be 6,010 sib pairs. Using the case-control model, only 820 affected individuals and controls are required to identify the same gene. (Calculations were performed as described in Ref. 12.) In fact, the higher statistical power of association studies has been theoretically calculated to allow even recognition of genes with relative risks as low as 1.5 using sample sizes that are reasonable owing to the ease of sample collection in this paradigm[8].

Do isolated populations offer an advantage in association analysis?

Although a consensus is emerging in the deliberation of linkage versus association studies, the possible advantages in the use of an inbred population originating from a limited number of founder individuals continue to be debated. Two advantages appear to be unequivocal. First, the use of homogeneous populations greatly reduces the chances of confounding because of population stratification, thus virtually eliminating the potential for artefactual associations caused by this factor[9]. Second, isolated populations exhibit reduced genetic heterogeneity, as demonstrated by a smaller number of polymorphic SNPs and by the significant reduction in the number of mutations found in specific disease-related genes. The reduced

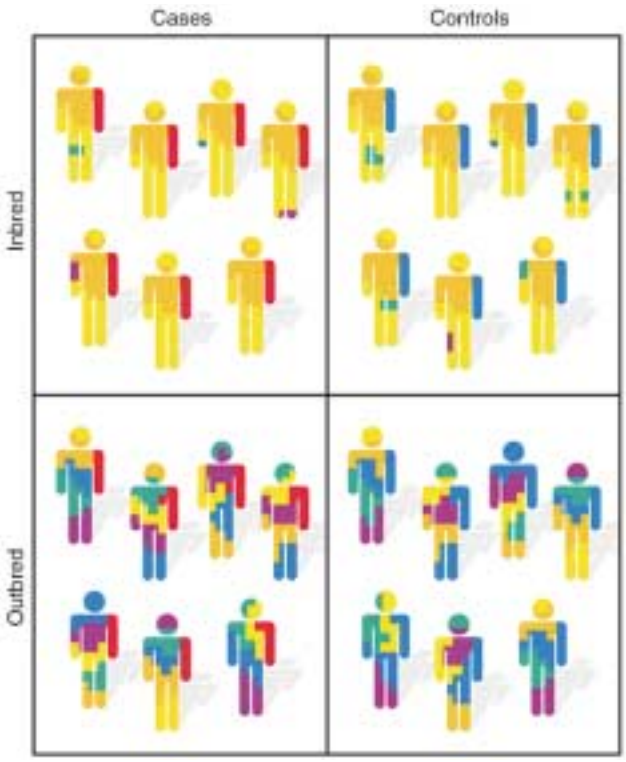

Figure 1. The signal to noise effect. Association scans are conceptually similar to the detection of a signal on the background of noise. In this Figure, the effect is graphically represented in the context of inbred versus outbred populations. When examining an inbred population it is immediately apparent that all the cases have a red right arm and all the controls have a blue right arm. Although this is also true in the outbred samples, it is much less obvious. Likewise in disease-gene discovery, the functional effect is similar in all populations, however the detection of a predisposing gene will be facilitated by examining the inbred population.

genetic heterogeneity has the effect of increasing the GRR because the similarity in genetic make-up decreases background noise[13]. (See Fig. 1 for a graphical explanation of this effect.)

The debate about the added advantage to be gained from the use of a founder population centres mainly on the question of whether these populations exhibit higher levels of LD (Refs 14,15). We have re-examined published data that was previously used to demonstrate that LD is only marginally improved in isolated populations, and have also simultaneously added our own data on Ashkenazi Jews, as an example of a highly homogeneous population[16,17]. In contrast to the previous analysis, which examined LD of all SNP pairs together, regardless of the distance between them, we differentiated between SNP pairs at distances below and over 200 kb. (The 200 kb threshold used was appropriate for the chromosomal region studied and may vary among different chromosomal regions). The rationale for this differentiation is that within short intervals factors other than distance (and the resulting recombination frequency) are more significant in influencing LD. These factors may include elements such as whether a new mutation occurs on the background of a common or rare haplotype[18]. However, because in isolated populations the anticipated higher level of LD is attributed to a smaller generation number and consequent reduced recombination, an increased level of LD is mainly expected between SNPs separated by a greater distance.

Indeed, no significant difference in the level of LD was found between SNPs at a distance of up to 200 kb when a heterogeneous CEPH group was compared with Finns, Sardinians or Ashkenazi Jews. For SNPs more than 200 kb apart, the level of LD for these three respective homogenous populations was increased by 4.7, 6.1 and 7.0 times, respectively, relative to the CEPH population[18]. Clearly, if these results demonstrated in a single genomic region are applicable to the entire genome, they indicate that to perform genomic scans with inbred populations will require significantly fewer markers than with outbred populations, and that there are differences in the level of LD even among various founder populations.

Most importantly, it must be recognized that an increase in GRR and an increase in LD both serve to decrease the theoretically required sample size for disease-gene detection independently from one another. Therefore their combined effect is multiplicative[18,19]. Thus even very small changes, when they occur simultaneously in these two factors, have very large effects on the required sample size. It is possible, for example, to compare the case of a hypothetical disease variant of a gene, which in a particular population has a GRR of 2 and an LD of 0.5, with the same gene on the background of an isolated gene population, in which its GRR is slightly increased to 3 and the LD in the chromosomal region in which it is found is increased to 0.75. In this instance the theoretically required sample size to detect the gene by the case-association paradigm is reduced from 4,151 affected and control individuals in the general population to 820 cases and controls in the founder population. (Calculations were performed as described in Ref. 12.)

Conclusion

In summary, the case-control model using an inbred, homogeneous population offers a feasible approach to chromosome scanning for disease genes, especially when decisions about marker density are also based on knowledge of local recombination frequencies. This approach for the discovery of disease-predisposing genes has strong statistical

power and minimizes the risk of false positives caused by population stratification. Following the initial detection of a disease-related gene or small chromosomal region in an isolated population, the results should be generalized, using an outbred population for confirmation.

A robust strategy for susceptibility gene discovery coupled with current developments in high-throughput genotyping technologies is expected to result in astounding progress in the dissection and understanding of complex diseases. This will lead to a totally new approach to the diagnostics and treatment of these common diseases and new challenges for the treating physician. Routine SNP genotyping of patients will provide more exact knowledge of disease subclasses, and the development of drugs targeted to different susceptibility genes will allow personalized, more effective and less hazardous treatment of disease.

References

1 Risch, N. (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* 46, 229–241
2 Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99
3 Carlson, C.S. *et al.* (2001) SNPing in the human genome. *Curr. Opin. Chem. Biol.* 5, 78–85
4 Shi, M.M. (2001) Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clin. Chem.* 47, 164–172
5 Concannon, P. *et al.* (1998) A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* 19, 292–296
6 Hutton, M. *et al.* (1998) Genetics of Alzheimer's disease. *Essays Biochem.* 33, 117–131
7 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
8 Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature* 405, 847–856
9 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181
10 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144
11 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204
12 Darvasi, A. *et al.* Power and efficiency of the TDT and the case-control design for association scans. *Behav. Genet.* (in press)
13 Khoury, M.J. *et al.* (1993) *Fundamentals of Genetic Epidemiology*, Oxford University Press, New York, USA
14 Eaves, I.A. *et al.* (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 25, 320–323
15 Taillon-Miller, P. *et al.* (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* 25, 324–328
16 Motulsky, A.G. (1995) Jewish diseases and origins. *Nat. Genet.* 9, 99–101
17 Risch, N. *et al.* (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. *Nat. Genet.* 9, 152–159
18 Shifman, S. and Darvasi, A. (2001) The value of isolated populations. *Nat. Genet.* 28, 309–310
19 Wright, A.F. *et al.* (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.* 23, 397–404