Mapping genes for complex traits in founder populations

C. OBER and N. J. COX*

Center for Medical Genetics, and *Department of Medicine, The University of Chicago, USA

Introduction

The advantages of inbred or founder populations for genetic studies were first noted over 30 years ago [1–3], but interest in these populations was renewed as a result of the explosion of molecular biological techniques and the initiation of the Human Genome Project [4,5]. The relatively small number of founders and recent ancestries that are characteristic of these populations facilitate the search for human disease genes and make them particularly amenable to novel analytic strategies. Founder populations have proven useful for mapping genes that underlie Mendelian disorders [e.g. 6–10] and multigenic disorders that segregate as Mendelian conditions in specific populations [11,12]. Particular features of population history that are unique to these populations (discussed below) have advanced the search for Mendelian conditions and should facilitate the search for genes that influence susceptibility to complex traits. This paper reviews the features of founder populations that make them useful for mapping studies and discusses the analytical challenges that founder populations pose to geneticists mapping complex traits.

The challenges of complex trait mapping

Identifying the loci that confer susceptibility to complex phenotypes, such as asthma, presents unique challenges to gene mappers. Unlike Mendelian traits, for which (usually) alleles at single loci determine the phenotype in a predictable manner, complex trait phenotypes are determined by alleles at many loci. Not only are the number of loci unknown, but the phenotypic affects of alleles at each susceptibility locus is presumably influenced by environmental factors. Furthermore, it is likely that many susceptibility alleles will be common polymorphisms and not ‘mutant’ alleles with obvious effects on the phenotype. Thus identifying genes that influence the expression of complex traits requires novel approaches and analytic strategies.

Two problems associated with mapping complex traits are particularly relevant to this discussion. First, it is assumed that complex phenotypes are genetically heterogeneous. That is, more than one genetic locus contributes to susceptibility and that the relative importance of alleles at different loci may vary from family to family. Second, the number of the susceptibility alleles with detectable effects are unknown, but for many complex diseases, such as asthma, it is assumed that many loci could potentially influence the phenotype. In addition, some proportion of these loci may have relatively modest effects on the phenotype.

In studies of outbred subjects, heterogeneity can be addressed by studying a very large numbers of families. This would increase the likelihood that within this heterogeneous population, a sufficient number of families will be sampled in whom the disease is influenced by the same locus. For example, if a gene A is important in only one-third of asthma families, one would want to sample enough families so that one-third of the sample would provide enough power to detect the effects of gene A. On the other hand, to identify susceptibility alleles with relatively modest effects on the phenotype requires using dense genetic maps as well as large numbers of families. This would ensure that any marker on the map will be sufficiently close to the gene to pick up these ‘modest’ effects in a linkage study.

Both sample size and marker density will have considerable effects on costs (Table 1). The cost estimates shown in Table 1 are based on a sample size of 500 subjects (or 125 sib pairs and their parents). Most genome-wide searches today study at least that many individuals and utilize a 10–12 cm map. But both denser maps and larger samples sizes would increase the likelihood of detecting all susceptibility loci. At the current time, however, this is usually not feasible.

The advantages of founder populations

Populations derived from a small number of founders offer theoretical advantages with respect to both required sample sizes and marker densities. First, the small number of founders decreases the number of independent susceptibility alleles segregating in the population (that is, it decreases heterogeneity). In these populations, there will be a relatively smaller number of loci influencing susceptibility to complex traits such as asthma; and it is theoretically possible that all affected individuals in the population will
have susceptibility due to the effects of the same alleles that are identical-by-descent (IBD) from a common ancestor.

Second, the relatively recent origins of the population ensures that the chromosomal region surrounding the disease allele that is IBD is larger than in the outbred population. As a result, less dense maps in these populations will provide similar information for linkage studies as more dense maps will provide in outbred populations. This is illustrated in Fig. 1. Two pairs of founder chromosomes are shown in the upper part of the figure. The black bar in the chromosome on the left represents a mutant allele. During meioses and the formation of gametes in each subsequent generation, recombination will shuffle the genetic material between homologous chromosomes. In the first generation, the offspring in this example inherited the disease allele on a chromosome with contributions from both of its parents’ chromosomes. The genetic material flanking the mutant allele will continue to be reshuffled in each subsequent generation, so that the segment of the chromosome around the mutation that is shared and ‘identical by descent’, or IBD, gets smaller in each generation. After many generations, as in outbred populations, the segment that is IBD will be quite small (<<1 cM). In populations with relatively recent ancestry, the size of the shared segment that is IBD will be much larger, and the actual size of this segment will be directly proportional to the number of generations since the founding. All alleles at loci within the shared segment that is IBD will have been preserved as a block from the original founder chromosome and will show linkage disequilibrium in the population. Therefore, in populations with a more recent common ancestry, loci with alleles in linkage disequilibrium will extend over a larger segment of the chromosome than in older populations. As a result, it will not only be possible to detect linkage over larger distances but it will also be possible to detect associations between alleles at linked loci over larger distances. Thus, linkage analyses in these populations can include shared chromosome segment analysis or allelic association studies, rather than, or as a supplement to more standard linkage analysis.

Table 1. Effect of sample size and marker density on cost. The sample size is equivalent to 125 sib pairs and parents

<table>
<thead>
<tr>
<th>Marker spacing (cM)</th>
<th>Number of markers per genome</th>
<th>Number of genotypes per 500 subjects</th>
<th>Cost at $2.50 per genotype $250,000</th>
<th>Cost at $1.00 per genotype $100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>200</td>
<td>100,000</td>
<td>$250,000</td>
<td>$100,000</td>
</tr>
<tr>
<td>12</td>
<td>350</td>
<td>175,000</td>
<td>$437,500</td>
<td>$175,000</td>
</tr>
<tr>
<td>5</td>
<td>3,000+</td>
<td>1,500,000</td>
<td>$3,750,000</td>
<td>$1,500,000</td>
</tr>
</tbody>
</table>

Fig. 1. Inheritance of a founder chromosome carrying a disease mutation. The mutation is designated by a black bar in one founder chromosome. Due to recombination between homologous chromosomes, the region flanking the mutation that is IBD will decrease in size each generation. See text for details.

Mapping strategies that take into account linkage disequilibrium between loci can be utilized most efficiently in founder populations, with varying degrees of resolution. The degree of resolution is directly proportional to the number of generations since the founding. This is illustrated in Table 2 for three founder populations in a hypothetical study of 50 individuals affected with the same autosomal recessive disease. The founders of the Finnish population lived ~80 generations ago. Linkage disequilibrium (or the size of shared chromosome segments that are IBD) should extend over ~1.25 cM throughout the genome. A chromosomal segment that is shared among the 50 affected individuals would be expected to be 0.025 cM, narrowing the gene search to ~25 kb of DNA. This region is small enough that it may contain as few as three genes. In the Hutterites, a population founded ~10 generations ago, linkage disequilibrium should extend on average over 10 cM, and shared segments that contain the gene of interest would be ~0.20 cM, or about 200 kb of DNA, a region large enough.
to contain up to 20 genes. In fact, empirical studies in the
Hutterites have shown that linkage disequilibrium does
indeed extend over 5–10 cM for most marker pairs in one
anonymous region of DNA [13]. In the population living
on the island of Tristan de Cunha [14], linkage disequi-
lbrium should be detected up to 17 cM, defining a region of
\( \approx 0.33 \text{ cM or 330 kb of DNA containing the disease locus.} \)
This region could contain as many as 33 genes. Thus, the
ease of detecting linkage or linkage disequilibrium is
inversely proportional to the ease of finding genes once
disequilibrium, or allelic association, is detected.

These estimations are based studying 50 individuals with
the same recessive condition. The predictions for complex
traits are not known. Nonetheless, the relationship between
the number of generations since the founding and the
resolution of association-based mapping approaches
should be the same. That is, the longer the time since the
founding, the more markers that will be needed to detect
associations. When associations are detected in older popu-
lations, the markers will be very close to the gene (i.e.
< 1 cM). In more recently founded populations, associations
will be detectable over greater distances and with less
dense maps, but these studies will provide poorer resolution
for finding the gene. Therefore, the most inclusive strategy
would be to study populations like Tristan or the Hutterites
for detecting linkage and linkage disequilibrium using less
dense maps (for example, 10–20 cM maps) and groups like
the Finns or even in an outbred population to narrow down
the region to a size that is amenable to physical mapping
(i.e. 0.5–1 cM). This strategy assumes that the locus asso-
ciated with the phenotype in the inbred population would
also be associated with the same phenotype in the more
outbred populations. Although there are no examples yet of
successes of this combined approach for mapping complex
traits, there is considerable overlap between the regions that
are showing evidence for linkage to asthma in the outbred
CSGA families [15] and inbred Hutterites [Ober and Cox,
unpublished data], indicating that this approach is at least
feasible.

### Analytical challenges of linkage studies in founder
populations

Linkage studies in founder populations pose a number of
challenges. Although the advantages of these populations
for mapping Mendelian conditions is widely appreciated,
the approaches used to assess the evidence for linkage to
complex phenotypes are not as straightforward. For exam-
ple, homozygosity mapping is an extremely useful approach
for mapping recessive conditions in inbred populations, but
has uncertain value in mapping genes for traits like asthma.
There is little (if any) evidence that asthma susceptibility
genes are recessive, and even in a population descended
from a limited number of founders there may be multiple
loci contributing to susceptibility to a disease as common as
asthma.

Traditional parametric approaches are problematic for
complex disorders in general [reviewed in 4], and have
additional complications for inbred genealogies. For exam-
ple, in the Hutterite pedigree there are literally hundreds of
inbreeding loops, and it is not possible to conduct analyses
with all the loops maintained. Breaking the loops requires
making many arbitrary decisions concerning the lines of
descent that are preserved and, inevitably, reduces the
power to detect linkage. Even after all loops have been
broken, the size of the Hutterite pedigrees precludes multi-
point linkage analyses unless the large, intact pedigree is
broken into smaller families. The size limitations of these
smaller families for multipoint analyses (e.g. in GENE-
HUNTER [16], 12–16 non-founders) insures that few
‘families’ will have a sufficient number of affected
individuals to be informative.

We have developed several strategies for linkage studies
of asthma in the Hutterites. For the initial marker screen,
the large pedigree was divided into three subpedigrees (each
about 150 individuals) and all inbreeding loops were
trimmed. We use a maximized likelihood in which the
penetrances, disease susceptibility allele frequency, and
recombination frequency are free parameters, although the

<table>
<thead>
<tr>
<th>Founder population</th>
<th>Number of generations since founding</th>
<th>Extent of linkage disequilibrium</th>
<th>Mapping resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland</td>
<td>80</td>
<td>1.25 cM</td>
<td>0.025 cM</td>
</tr>
<tr>
<td>Hutterites</td>
<td>10</td>
<td>10 cM</td>
<td>0.20 cM</td>
</tr>
<tr>
<td>Tristan de Cunha</td>
<td>6</td>
<td>17 cM</td>
<td>0.33 cM</td>
</tr>
</tbody>
</table>

© 1998 Blackwell Science Ltd, Clinical and Experimental Allergy, 28, Supplement 1, 101–105
penetrances and allele frequencies are constrained by the population prevalence of disease. The maximized likelihood is compared to one in which the penetrances and disease susceptibility allele frequency are free, but the recombination frequency is constrained to be 0.5 \cite{17}. A likelihood ratio $\chi^2$ with 1 degree of freedom can be used to assess the evidence for linkage. The advantage of this approach is that a mode of inheritance for asthma does not need to be specified and in this sense the test is ‘semiparametric’. However, because the inbreeding information is not included, it is likely that considerable power is sacrificed.

Less traditional approaches for assessing the evidence for linkage, such as shared segment analysis, are clearly attractive alternatives in founder populations. However, even these approaches have their limitations in the Hutterite genealogy. In particular, it is impossible to know a priori how much sharing is expected among affected individuals because all individuals in the pedigree are related to each other through a number of different paths. For example, the average degree of relationship between siblings in an outbred family is 0.5 (i.e. among siblings, 50% of genes on average are IBD from one parent). Among Hutterites, the average degree of relationship among siblings is >> 0.5 because the parents of the sibship are related to each other through many paths. Thus, unless all paths connecting all pairs of individuals are taken into account, the expected proportion of allele sharing between affected individuals will be underestimated and the significance attributed to the observed sharing will be overestimated.

A more appropriate, and potentially more powerful approach for assessing linkage in the Hutterites would be to compare the observed frequency of segment or allele sharing among affected individuals to the expected frequency of sharing based on the genealogical relationships between these individuals. However, it is not at all trivial to estimate the expected amount of sharing in a pedigree of this complexity, and this approach would be quite sensitive to misspecifications of relationships in the genealogy. For example, incorrect relationships between Hutterites in the first few generations of the pedigree could have quite profound effects on the estimates of expected sharing among individuals sampled from the most recent generations. Deviations between observations and expectations could result from unknown relationships among the founders, which would increase the genetic similarity that was observed compared to expectations based on the existing genealogy. On the other hand, mistakes in the genealogies, non-paternity, and unrecorded adoptions might decrease the observed genetic similarity compared with expectations. To assess the magnitude of this potential problem, we have initiated a study to assess the fit of the genotypic data derived from a complete genome scan to the existing genealogical information in the Hutterites. Based on the results of this study it will be possible to determine whether expectations of sharing derived from the genealogy are adequate for linkage studies, whether the genealogy can be ‘fixed’ to better fit the genotypic data, or whether the observed genotypic data averaged over all loci provides a more appropriate estimate of the null expectation for sharing among affected individuals.

It is likely that the true power of founder populations for mapping studies of complex phenotypes will depend on our ability to assess the evidence for increased allele sharing or segment sharing among affected individuals by taking into account all known genealogical relationships as well as the background level of allele or segment sharing averaged over all loci in a genome-wide screen. This approach would take into account both the genealogical and genomic data and provide the most robust expectations for linkage studies.

Conclusions

Founder populations offer considerable advantages for mapping complex traits because of their recent ancestries and small number of founding genomes. In theory, linkage studies in founder populations should require less dense maps and smaller sample sizes than similar studies conducted in outbred populations. In addition, founder populations are more amenable to novel analytic strategies, such as linkage disequilibrium mapping and shared segment analysis. However, the large, complex genealogies that are characteristic of these populations pose considerable analytical and computational problems and interpretation of linkage results are not always straightforward. Nonetheless, the development of new approaches that take into account both genealogical relationships and empirical estimates of allele sharing should be quite powerful for assessing the evidence for linkage to complex phenotypes in these populations.

Acknowledgements

This work was supported by NIH grants HL 49596 and HL 56399.

References
