The value of isolated populations

The relative advantages of isolated (also termed founder or homogeneous) and outbred populations in identifying genes affecting complex traits has been the subject of much debate[1–4]. The extent of linkage disequilibrium (LD) in isolated and outbred populations has been charted, and the small difference between the two types of population has been taken to indicate that the advantage of isolated populations is possibly minimal[3,4].

Here, we argue that the usefulness of isolated populations in the dissection of complex traits should not be dismissed. First, the apparent similarity of LD between isolated and outbred populations does not consistently hold, particularly for single-nucleotide polymorphisms (SNPs) relatively far apart from one another. Second, isolated populations usually have reduced genetic heterogeneity, which can significantly increase genotypic relative risk (GRR) and hence the ability to identify genes.

The measured LD level between two SNPs is determined by four elements: (i) the distance between the two SNPs; (ii) the age of the younger SNP; (iii) the frequency of the SNP allele adjacent to the new mutation (Fig. 1); and (iv) probabilistic processes of recombination, mating, drift, population growth, bottlenecks, and so on.

As often observed[5–7], LD decays as distance increases. It is, however, important to notice that, within short intervals, distance itself does not significantly influence the expected level of LD. The reason might be that, within this range, the reduction in LD, as a function of distance, is negligible compared with the other three elements (ii–iv). For example, the context of a new mutation—whether it occurs on a common or rare haplotype—can have a highly significant effect on LD (Fig. 1). As distance increases, however, the increased recombination frequency will weaken any association, undermining the effect of the elements ii–iv. In isolated populations, a higher level of LD is attributed to the smaller generation number, which results in reduced recombination. By this logic, an increased level of LD in an isolated population is mainly expected between SNPs separated by a greater distance; here, recombination is the primary element that determines the level of LD.

In light of this, we have re-examined (Fig. 2) the results presented in Table 2 of Taillon-Miller *et al.*[4], separately analyzing the average LD for SNPs at distances below and over 200 kb (the 200 kb threshold may vary and is appropriate only for this chromosomal region). We have also analyzed another isolated population, Ashkenazi Jews, to further support the generality of the findings. For the Ashkenazi population, 5 of the 15 markers in Table 2 of Taillon-Miller *et al.*[4] were not polymorphic or were not amplified.

For each pair of SNPs, we calculated the ratio: r^2 in an isolated population over r^2 in the CEPH samples (pairs of SNPs with $r^2=0$ were not included in the analysis). We presented the average of these ratios over all pairs separately for the Finnish, Sardinian and Ashkenazi populations for SNPs at distances below and over 200 kb (Fig. 2). As anticipated and previously reported, there is no significant difference in the level of LD between SNPs at a distance of up to 200 kb between isolated and outbred populations: the average of the r^2 ratios was close to 1. When, however, the distance between SNPs is greater than 200 kb, the level of LD (represented by the r^2 ratios) increases by 4.7, 6.1 and 7.0 (average 5.6; $P=0.002$) times for the Finnish, Sardinian and Ashkenazi samples, respectively.

If these results are applicable to the entire genome, they indicate that to carry out genome scans with outbred populations one will require significantly more markers or a significantly larger sample size than with isolated populations. Note, for example, that the required sample size when testing a single, specific SNP is proportional to the r^2 between that SNP and the functional polymorphism: a sixfold increase in r^2 means a sixfold increase in the sample size required to identify the specific gene in question.

It has been predicted that LD of a useful magnitude will not extend beyond 3 kb and that isolated populations will not provide any benefit unless it is the product of a very narrow bottleneck or the SNP in question is of low frequency (<5%)[8]. Our results and those of others[4,7] do not, however, support this expection. Collins *et al.*[7] have suggested that the discrepancy between expectation and empirical results is owing to the fact that the expectation is based on the assumption of monotonic population expansion. In reality, however, populations probably undergo significant cycles of expansion and bottlenecks.

Genetic heterogeneity can dramatically affect the ease (or difficulty!) with which complex traits are dissected. Allelic heterogeneity is the presence of a number of SNPs within a given gene, each affecting susceptibility to disease. Nonallelic heterogeneity is the presence of SNPs in different genes, each affecting susceptibility to disease[9].

It is reasonable to assume that isolated populations have less genetic heterogeneity. This is exemplified by the reduced number of polymorphic SNPs observed in several studies (for example, the percentage of nonpolymorphic SNPs is 13% and 56% for the Finns and Sardinians respectively[4]). Reduction in genetic heterogeneity is also illustrated by the significant reduction in the number of mutations found in specific disease-related genes, for example *BRCA1* and *BRCA2* in Ashkenazi Jews[10]. Animal models suggest that a large number of genes affect complex traits[11]. Thus, in homogeneous populations, the anticipated absence of polymorphisms in a subset of the relevant loci may reduce genetic variation and increase statistical power for the identification of genes.

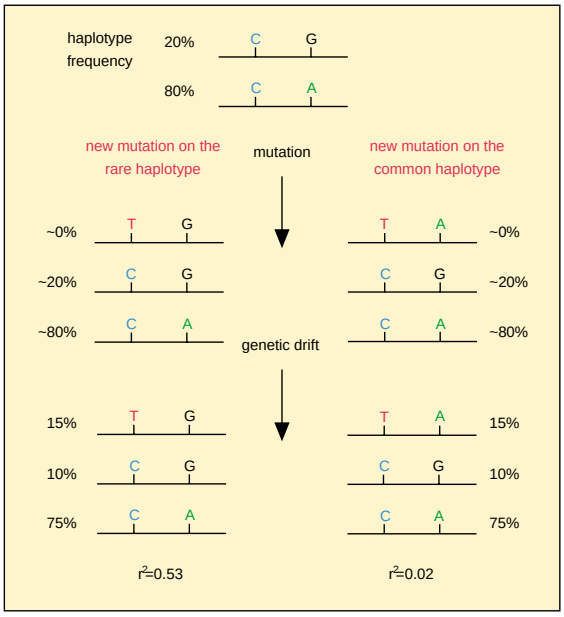

Fig. 1 The effect of phase on LD. The two SNPs are closely linked so that effectively no recombination has occurred between them since the creation of the younger SNP. The level of LD is strongly affected by the frequency of the haplotype on which background the younger mutation originally occurred. LD levels are represented by r^2; between the two cases, there is a 25-fold difference. Phase has a greater effect than distance on LD between close SNPs.

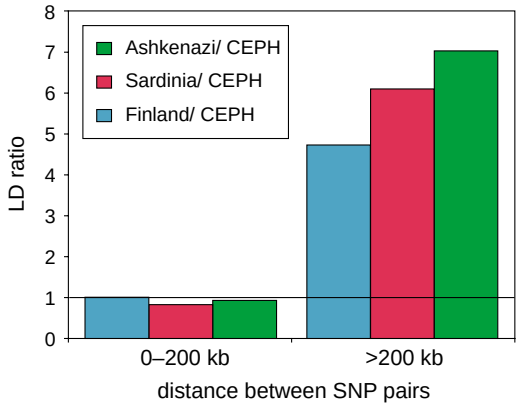

Fig 2 We calculated the ratio of r^2 for each pair of SNPs in samples of isolated and outbred populations. The average of these ratios is plotted separately for the Finnish, Sardinian and Ashkenazi samples, and for pairs at distances exceeding or less than 200 kb. For the Finnish/CEPH and Sardinian/CEPH, we made 18 and 21 comparisons for distances exceeding and less than 200 kb, respectively; for Ashkenazi/CEPH populations, we made 12 and 5 comparisons. The sample sizes of CEPH, Finnish, Sardinian and Ashkenazi populations were 92, 100, 150 and 100, respectively.

Additional work is necessary to estimate with accuracy the advantage of an isolated population owing to reduction in genetic heterogeneity. It is, however, worth noting that reduction in genetic heterogeneity directly affects the GRR, which is roughly proportional to the square root of the sample size required to identify a given gene[9]. Consequently, even a moderate effect on the GRR may have a significant effect on the required sample size.

The differences in LD observed between isolated and outbred populations on one hand, and the expected difference in the extent of genetic heterogeneity on the other, are independent and thus have a multiplicative effect on the required sample size. For example, if, because of differences in LD, the sample size needs to be increased threefold, and if, because of the differences in GRR another threefold increase is required, the sample size will need to be increased ninefold with outbred populations to achieve power comparable to that of an isolated population.

An additional advantage of using isolated populations is the applicability of the case-control design. The case-control design promises to be the most efficient paradigm for the association of genes with complex phenotypes[9]. One of the disadvantages of the case-control paradigm is its susceptibility to false-positive results through population stratification, if the cases and the controls are not well matched. With isolated populations, this is less of a risk because of the known and homogeneous genetic background of all individuals.

Sagiv Shifman[1] & Ariel Darvasi[1,2]
[1]Department of Evolution Systematics and Ecology, The Hebrew University of Jerusalem, Jerusalem 91904, Israel. [2]IDgene Pharmaceuticals Ltd, Beit Ofer – 5 Heftzadi Street, PO Box 34478, Jerusalem 91344, Israel. Correspondence should be addressed to A.D. (e-mail: ariel@idgene.com).

1. Abbott, A. *Nature* **406**, 340–342 (2000).
2. Boehnke, M. *Nature Genet.* **25**, 246–247 (2000).
3. Eaves, I.A. *et al. Nature Genet.* **25**, 320–323 (2000).
4. Taillon-Miller, P. *et al. Nature Genet.* **25**, 324–328 (2000).
5. Jorde, L.B. *Am. J. Hum. Genet.* **56**, 11–14 (1995).
6. Ott, J. *Proc. Natl. Acad. Sci. USA* **97**, 2–3 (2000).
7. Collins, A., Lonjou, C. & Morton, N.E. *Proc. Natl. Acad. Sci. USA* **96**, 15173–15177 (1999).
8. Kruglyak, L. *Nature Genet.* **22**, 139–44 (1999).
9. Risch, N.J. *Nature* **405**, 847–856 (2000).
10. Roa, B.B., Boyd, A.A., Volcik, K. & Richards, C.S. *Nature Genet.* **14**, 185–187 (1996).
11. Frankel, W.N. *Trends Genet.* **11**, 471–477 (1995).

DelBank: a mouse ES-cell resource for generating deletions

Chromosomal deletions are valuable reagents for identifying and mapping genetic function. To simplify the creation of deletions in mice, we developed a collection of embryonic stem (ES) cell clones called DelBank. DelBank is based upon a technology in which radiation-induced deletions of herpes simplex virus thymidine kinase (*tk*) cassettes, inserted by homologous recombination into F_1 hybrid ES cells, are selected with the antiviral drug 1-2'-deoxy-2'-fluoro-β-D-arabinofuranosyl-5-iodouracil (FIAU)[1,2]. The deletions range in size from less than 1 cM to over 20 cM. Cells of the F_1 generation retain germline competence following irradiation, and the heterozygosity enables the molecular mapping of deletion endpoints.

DelBank consists of more than 90 F_1 hybrid ES cell clones, each containing a randomly integrated *tk*-containing vector. We generated clones using either a plasmid vector (pBanTKcass[1]; Fig. 1*a*) or modified retroviral poly(A) trap vector[3] (Fig. 1*b*). These were introduced into v6.4 (C57BL/6J×129S4/SvJae) or v17.2 (BALB/cJ×129S4/SvJae) ES cells[1,4] by electroporation (plasmid) or retroviral infection (gene trap), followed by selection for

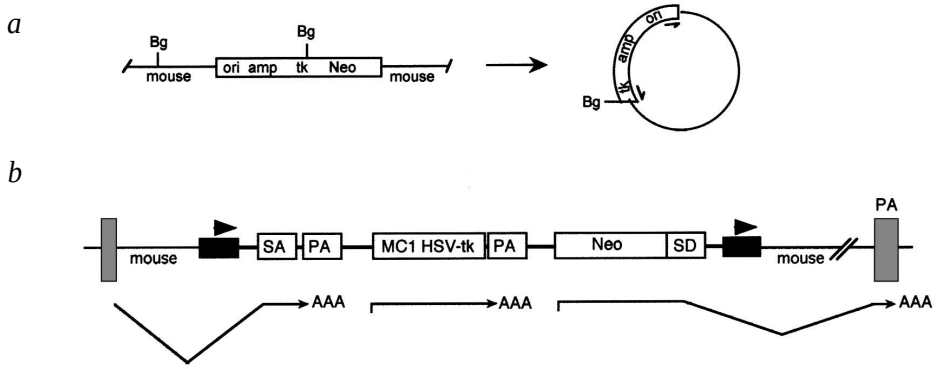

Fig. 1 DelBank vector strategies. ***a***, Plasmid rescue. DNA from transformed ES cell clones was cleaved with *Bgl*II, circularized and transformed into bacteria. The sequence of the 'rescued' mouse DNA was determined to generate amplimers that identify SSCPs. ***b***, Gene trapping. PA, poly(A) signal; SD, splice donor. Shown is a productive gene trap, where transcripts from an endogenous gene (left) undergo splicing into a vector splice acceptor (SA) and are terminated at vector PA, while the neo gene acquires a host PA. Arrowheads indicate LoxP sites.