Mining population-level DNA repositories and SNP databases

Joel S Bader
Director of Bioinformatics, CuraGen Corp, USA

Traditional linkage-based studies of human genetics have revealed the genetic basis of uncommon hereditary diseases but have stalled in uncovering the genetic variants that predispose individuals to complex diseases such as cancer, diabetes, cardiovascular disease, and psychiatric disorders. Population-level DNA repositories, SNP databases, and high-throughput genotyping may provide the raw material needed to make these discoveries.

The old and the new

Newer population-level studies differ from traditional family-based studies. To help advance our disease programs, we are shaping CuraGen's human genetics infrastructure to account for these differences. We discuss their impact on the three principal components of a human genetics study: the disease or trait under investigation, the population in which the trait is measured and whose DNA is collected, and the genetic markers used to identify genes that contribute to disease risk.

Diseases

Two broad categories of human traits are shown in Figure 1A. The left panel shows a simple trait that partitions a population into distinct sub-populations, in this case a larger unaffected fraction and a smaller affected fraction. Clear boundaries between sub-populations are typical of a monogenic disorder: a single gene at a control point in a biological process has an allelic variant with an aberrant mutation, and individuals who inherit the allele (one copy for a dominant allele, two copies for a recessive allele) exhibit the disease.

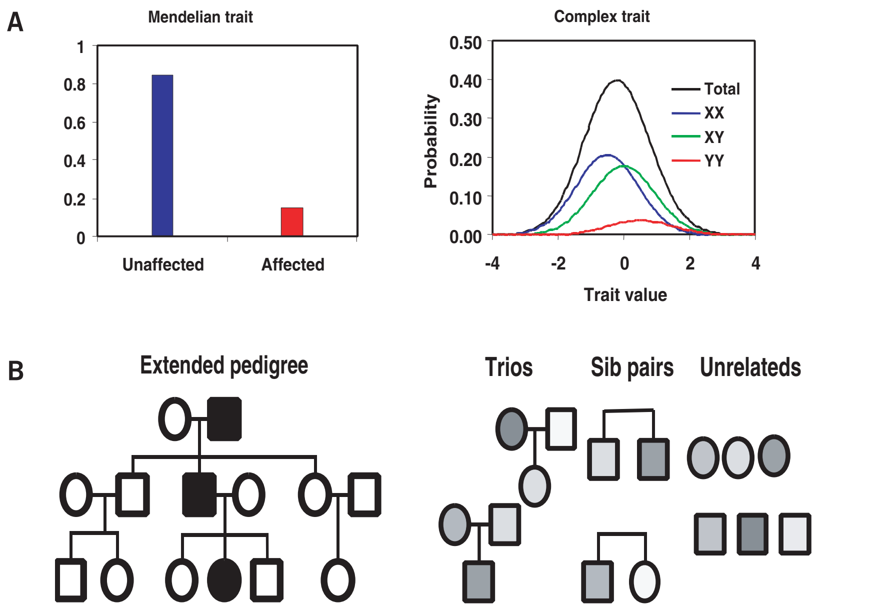

Figure 1. Phenotypes and family structures examined using traditional genetic linkage (left) contrast with newer population-level association tests (right). (A) Linkage tests are designed for traits with Mendelian inheritance, while association tests are more powerful for quantitative traits more characteristic of complex disease. The quantitative trait population has been stratified according to genotype at a biallelic locus (alleles X, Y) for a functional polymorphism. (B) Extended pedigrees with multiple affected individuals are informative for linkage tests, while trios, sib pairs, and unrelateds are informative for association and are easier to collect. Mendelian traits permit a dichotomous classification, while quantitative phenotypes are indicated by level of shading.

Complex diseases, such as obesity and hypertension, lack a clear distinction between individuals who are unaffected or affected. The values for traits underlying these complex diseases often follow bell-shaped normal distributions, as shown in the top right panel. These types of distributions arise automatically when multiple genetic and environmental factors contribute to a trait.

Variations in several genes, termed quantitative trait loci (QTLs), may influence the trait value, each adding its own small shift. Variations at several independent loci may be required to push an individual into a tail of the distribution. The tail may be defined precisely by introducing a threshold value; individuals may be classified as affected (above threshold) or unaffected (below threshold). This classification may serve more as a clinical convenience than as a real distinction.

The quantitative measure necessarily carries more information than the derived classification. In Figure 2 we show how the fraction of information retained by a threshold-based dichotomous classification depends on the classification threshold. The number of individuals required for a study is inversely proportional to this information measure. As is seen from the figure, changing a quantitative trait to a qualitative classification can drastically increase the population size required for a study.

Populations

If a disease is monogenic with Mendelian inheritance, some families may have multiple affected individuals across several generations. These families are informative: affected individuals in a single pedigree are likely to have

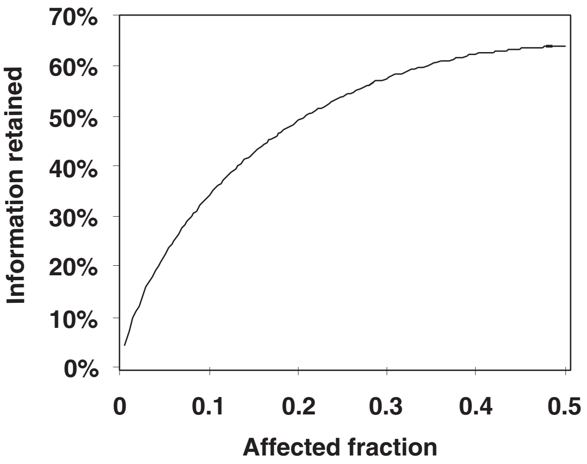

Figure 2. The information retained after a dichotomous affected/unaffected classification based on an underlying quantitative phenotype is shown as a function of the fraction of the population classified as affected. If the most extreme 10% of individuals are classified as affected, for example, 34% of the information is retained, and the population required for a study based on the dichotomous classification is 3-fold larger than for a study based on the underlying quantitative phenotype. See Bader *et al* (2001) for details.

inherited the same aberrant variant, and identifying a chromosomal region that they all share provides evidence that the disease gene is linked to this location. The regions that can be identified by a traditional linkage study are about 106 nucleotides (nt), roughly the sizes of the physical chunks of DNA that recombine from parentally-inherited chromosomes to make gametes.

Pedigree-based studies are not as useful for complex diseases: the disease is not inherited as a simple Mendelian trait, and individuals who inherit an aberrant variant, pushing their trait value to the tail of the distribution, may also inherit a protective variant that returns them to the center. Furthermore, environmental effects, such as changes in nutrition or public health services, may confound cross-generational comparisons. Instead of traditional linkage, scientists look for direct association of a particular gene variant with a trait. This variant may be the functional variation itself or may be linked to the functional variation at the population level.

Since the effect from any single genetic polymorphism is likely to be small, large populations are required to measure an effect with statistical significance. These needs have spurred the establishment of DNA repositories as national and commercial resources. For quantitative traits, individuals contributing to these repositories may be recruited from the general population; disease-specific populations are not necessary. Common family structures include parent-child triads, siblings, and unrelated individuals. Populations for traditional linkage studies and newer association studies are compared in Figure 1B.

Markers

Markers for traditional linkage studies are spaced commensurate with single-generation recombination distances, 106 to 107 nt as mentioned previously, with several hundred markers required for a full genome scan. For association studies, the marker spacing must be commensurate with linkage disequilibrium at the population level, 10,000 to 100,000 nt, and thousands to hundreds of thousands of markers are required. Single-nucleotide polymorphisms (SNPs) provide a suitably dense marker set.

While a greater number of markers implies a greater genotyping cost, it also carries the benefit that a finer map yields a much smaller candidate region. At CuraGen, we enhance the benefits of SNP-based markers by focusing on polymorphisms that occur in pharmaceutically tractable genes coding for proteins that may serve as therapeutics, antibody targets, and small molecule targets. We preferentially select SNPs that cause amino-acid changes and that may affect transcriptional or translational regulation.

These gene-based SNP markers, derived in part from proprietary cDNA sequences, may actually be the sought-after functional polymorphisms underlying disease risk and are likely to identify proteins that can enter a drug development pipeline. In contrast, SNP markers from non-coding genomic regions are unlikely to affect biological processes directly and are less likely to yield tractable targets.

Current issues in human population genetics

SNPs versus haplotypes

In traditional studies to detect linkage, single-point tests often have lower power

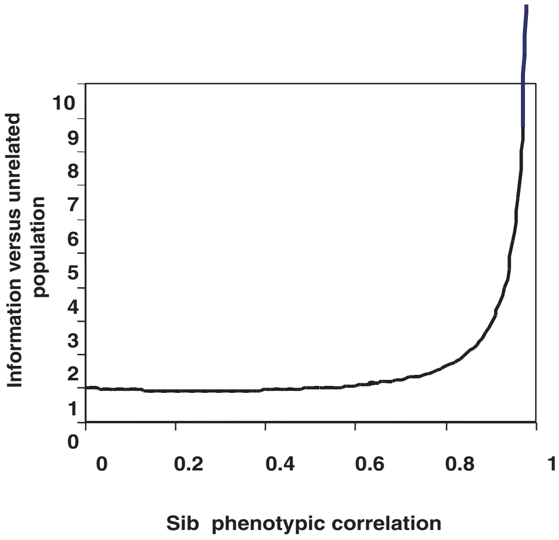

Figure 3. The information provided by a population of sib pairs is shown relative to the information from a population of unrelated individuals. With a 75% phenotypic correlation, for example, the sib pair population is 1.4-fold more efficient, and a study using sib pairs requires only 70% of the population that would be required for a study using unrelated individuals.

AD

than multi-point tests. For association studies using SNP markers, we found the opposite result: single-point tests based on genotypes at each SNP locus may have greater power than multi-point tests based on the observed haplotypes. Our results were based on candidate gene studies in which it is reasonable to assume that the functional mutation is in the marker set.

Suppose, for example, that certain individuals do not respond to a particular drug. At the start of a pharmacogenetics study, it is straightforward to identify the SNPs that occur in the drug target: sequencing 100 chromosomes is sufficient to identify 95% of the variants that occur at 3% allele frequency or above. If a functional polymorphism exists, it is likely to be found.

When a SNP-based marker set does not include the functional polymorphism, for example in a whole-genome scan, haplotype-based tests may provide benefits. Nevertheless, SNP-based tests have the additional benefit of permitting cost-saving pooled DNA tests (see below) that are not yet available for haplotypes.

Population stratification

With population stratification, also termed admixture, a heterogeneous population can be partitioned into more homogeneous sub-populations, often by ethnicity. Stratification degrades the power of population-level tests. If environmental factors such as diet or smoking are an important factor in disease risk, for example, and these risk factors correlate with ethnicity, then genetic variants that also correlate with ethnicity may be identified spuriously by population-level association. An increase in the overall phenotypic variance in a heterogeneous population may lead to false negatives as well.

One approach to stratification is to test each marker for association within homogeneous sub-populations defined statistically through the remaining markers. Early results are promising, suggesting that population stratification may no longer be an impediment to population-level studies.

A second solution is to apply within-family tests, which are generally immune to stratification effects. If measurements from siblings are available, each may be used as a control for the other. Even more desirable are twin-based sample collections due to greater similarity of environmental and age-related factors. The power of twin-based studies motivated CuraGen's collaboration with Gemini Genomics, a leader in twin-based sample collections.

Even in the absence of stratification, sibling collections may be more powerful than collections of unrelated individuals. Figure 3 shows the information contained in a population of sib pairs relative to the same number of unrelated individuals. The sib population is roughly equivalent when the trait correlation between sibs is low to moderate and has a clear advantage when the correlation is high. This finding runs counter to expectations based on dichotomous traits that unrelated populations are usually more powerful. For reference, sib correlations for complex diseases due to shared genetic and environmental factors range from 25% to 75%.

Cost reduction through pooling

In traditional genetic studies, each individual is genotyped at every marker. This design can be prohibitively expensive for population-level studies involving thousands of markers. Tests of pooled DNA can provide tremendous savings: individuals with extreme high and low phenotypic values are selected, their DNA is pooled, and the allele frequency of each marker is determined for each pool. A frequency difference between pools indicates association.

We recently reported optimal selection criteria for pooled tests using human DNA repositories. Neglecting experimental error, pooling the high and low 27% tails of a population maximizes the test power and contains 80% of the information provided by full genotyping. Experimental error in pooled measurements, typically (1-2% raw allele frequency, decreases the information provided, but enough information remains to allow pooled tests to be used as a pre-screen with associations confirmed by individual genotyping.

Conclusion

With the rising availability of population-level DNA databases and SNP markers, association studies have the potential to identify the genetic variants underlying complex disease. Intuition based on expectations from traditional linkage analysis of dichotomous Mendelian traits may be inaccurate when applied to population-level association studies of complex, quantitative traits. A driving force behind population-level studies may be the development of inexpensive SNP genotyping assays, in particular allele frequency measurements of pooled DNA.

Joel S Bader
Director of Bioinformatics
CuraGen Corp
555 Long Wharf Drive
New Haven
CT 06511
USA

Email: jsbader@curagen.com

FURTHER READING

Association tests
Risch NJ & Merikangas K (1996) **The future of genetic studies of complex human diseases.** *Science* **273**:1516-1517.

SNPs versus haplotypes
Bader JS (2000). **The relative power of SNPs and haplotypes as genetic markers for association tests.** *Pharmacogenomics* **2**:11-24.

Population stratification
Bacanu SA *et al* (2000) **The power of genomic controls.** *Am J Hum Gen* **66**:1933-1944.

Pritchard JK *et al* (2000) **Association mapping in structured populations.** *Am J Hum Gen* **67**:170-181.

Zhang S & Zhao H (2001) **Quantitative similarity-based association tests using population samples.** Manuscript submitted

DNA pooling
Bader JS *et al* (2001) **Efficient SNP-based tests of association for quantitative phonotypes using pooled DNA.** *Genescreen* in press.