

species. For example, repeated sequences have been shown to act as hot spots for recombination events in previously studied genomic systems^{4,5}. Indeed, some of the conserved segments in *Mycoplasma genitalium* and *Mycoplasma pneumoniae* are surrounded by repeated sequences⁶.

Dreams, dollars and microbial evolution

Comparative genomics should now become an object of intense study. This young field of research has not only triggered a heated debate about the universal tree of life, but also revealed spectacular details of molecular evolution. As shown by Tillier and Collins, a comparison of the two *Chlamydia* genomes have provided indications for replication-directed translocation². Likewise, comparative analyses of Rickettsial and Mycobacterial genomes have disclosed the secrets of genome degradation and DNA elimination^{7,8}. The genetic mechanisms whereby novel gene functions are created by the inflow of external DNA are currently being intensively studied in model organisms such as *Escherichia coli* and its relatives⁹. We all know that horizontal gene transfers occur—occasionally—but do they really occur at frequencies high enough to

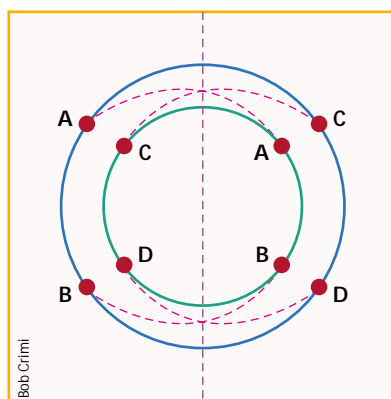


Fig. 2 Schematic illustration of gene order structures (A, B, C, D) that are mirror-imaged across the replication axes in two genomic sisters, possibly due to translocations at the replication forks.

influence our basic understanding of how living organisms are related to each other¹⁰?

The current findings underscore the fact that a great deal of work is yet required to quantify the rates at which rearrangement processes occur in bacteria. So far, microbial species have mostly been selected for genome projects based on their practical value in medicine and biotechnology. By coincidence, some of these projects

(including the study by Tillier and Collins²) have also yielded information about microbial evolution. In my view, however, microbial genome projects should be selected based on evolutionary criteria and in a more systematic manner. In combination with the development of appropriate mathematical tools for complex combinatorial analyses, it should be possible to generate evolutionary models that accurately describe the phylogenetic relationships of micro-organisms based on their genomes. With the help of such models, phylogenetic schemes may eventually be defined that incorporate all of the various processes whereby microbial genomes evolve. □

1. Kimura, M. *Nature* **217**, 624–626 (1968).
2. Tillier, E.R.M. & Collins, R.A. *Nature Genet.* **26**, 195–197 (2000).
3. Read, T.D. *et al.* *Nucleic Acids Res.* **238**, 1397–1406 (2000).
4. Krawiec, S. & Riley, M. *Microbiol. Rev.* **54**, 502–598 (1990).
5. Segall, A., Mahan, M. & Roth, J.R. *Science* **270**, 397–403 (1988).
6. Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Hermann, R. *Nucleic Acids Res.* **25**, 701–725 (1997).
7. Andersson, J.O. & Andersson, S.G.E. *Mol. Biol. Evol.* **16**, 1178–1191 (1999).
8. Andersson, J.O. & Andersson, S.G.E. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
9. Ochman, H., Lawrence, J.G. & Groisman, E.A. *Nature* **405**, 299–304 (2000).
10. Doolittle, W.F. *Science* **284**, 2124–2128 (1999).

Guilt by association

David Altshuler^{1–3}, Mark Daly³ & Leonid Kruglyak^{4,5}

¹Departments of Genetics and Medicine, Harvard Medical School, Boston, Massachusetts, USA. ²Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. ³Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts, USA.

⁴Howard Hughes Medical Institute, ⁵Division of Human Biology, Program in Genetics, Fred Hutchinson Cancer Center, Seattle, Washington, USA.
e-mail: altshul@genome.wi.mit.edu, mjdaly@genome.wi.mit.edu & leonid@fhccr.org

Positional cloning of common disease genes is a central but elusive goal of human geneticists. Progress is now reported by Bell and colleagues in their study of NIDDM1, a locus implicated in type 2 diabetes. The complex nature of the reported association illustrates the challenge of implicating a specific gene and mutation in the causation of polygenic disease.

In the United States criminal justice system, guilt must be proven “beyond a reasonable doubt,” a stringent standard reflecting the serious consequences that follow a guilty verdict. Perhaps because of these high stakes, ‘proof’ is frequently a subject of heated debate. With positional cloning, similar issues arise when the ‘guilt’ of a specific gene suspected of disease causation must be proven. The stakes are high because of the intense interest ‘disease’ genes elicit in patients, clinicians and biologists alike. Proving causation, however, is not necessarily a straightforward proposition, even for relatively simple cases of mendelian diseases. In some instances, a ‘smoking gun’ may be

found in the form of different disrupting mutations in a single gene. The case for the prosecution is more difficult when only missense changes or alterations in non-coding regions are observed. Nonetheless, culprits in well over one hundred mendelian diseases have been snared by positional cloning.

For common diseases with complex inheritance, convicting a guilty party poses a far greater challenge. Genome-wide linkage scans designed to localize disease genes have yielded few significant findings, and failure to reproduce published linkage results is endemic. The lack of strong linkage signals indicates that, for most complex traits, no single locus con-

fers a high degree of risk. Because mutation of any single gene is neither necessary nor sufficient, the correlation between genotype and phenotype is imperfect, and individual recombinants cannot be trusted for fine-mapping. This means that implicated genomic intervals are large, and yet do not include the gene with 100% confidence^{1,2}. Finally, as mutations contributing to complex traits may be subtle alterations of gene function or regulation, rather than disruptions of coding sequence, they may be difficult, if not impossible, to recognize from primary sequence data alone. For all these reasons, there are no examples of ‘complex disease’

genes cloned purely by position (aside from highly familial subtypes).

On page 163 of this issue, Yukio Horikawa, Naohisa Oda, Nancy Cox, Graeme Bell and colleagues report on their pioneering effort to positionally clone a gene that affects susceptibility to type 2 diabetes in Mexican Americans. Through a combination of clever detective work and brute force, they indict a gene encoding calpain-10 (*CAPN10*), and show that certain combinations of polymorphisms in this gene are associated with the risk of type 2 diabetes. By indicating a role for a calpain protease, these findings propose a fundamentally new hypothesis for diabetes research. But because no smoking gun is found, it is a great challenge to convict the guilty party (that is, gene and mutation) beyond a reasonable doubt. The study provides an unprecedented look at the cutting edge of positional cloning for complex traits and teaches lessons of lasting importance about just how difficult finding genes for common diseases may turn out to be.

A dragnet for suspects

In 1996, Hanis *et al.* reported significant evidence for linkage of type 2 diabetes to the distal long arm of human chromosome 2 (the locus was designated *NIDDM1*; ref. 3). The implicated region was large, with the 1-*lod* support interval, which is expected to contain the responsible gene in 80–90% of cases¹, extending over 12 cM. The discovery of an interaction between *NIDDM1* and a putative locus on chromosome 15 (ref. 4) focused attention on a sub-region of 7 cM. At this point, good fortune intervened—construction of a physical map revealed that the 7 cM spans only 1.7 million base pairs, rather than the expected 7 million (the average ratio of physical distance to genetic distance across the human genome is approximately 1 million base pairs per cM). This unusually high recombination rate is probably a consequence of the telomeric location of *NIDDM1*.

But finding the causative gene(s) in 1.7 Mb of sequence is a formidable task that has no generic solution. This particular interval contains at least 7 known genes and 15 expressed sequence tags (ESTs): none of these are obvious candidates. Horikawa *et al.* chose to screen polymorphisms in the region for association with diabetes, relying on linkage disequilibrium (LD) to lead them toward their quarry. An initial examination of 21 SNPs

detected a hint of association between diabetes and multi-locus haplotypes at 3 consecutive SNPs. This haplotype frequency difference prompted the identification and typing of additional SNPs in the region, some of which were nominally associated to diabetes. The authors then sequenced a 66-kb interval in the region (centred on the associated SNPs) in 10 diabetics, revealing 3 genes (including *CAPN10*) and 179 polymorphisms.

phism within *CAPN10*. The common G allele (present on 75% of normal chromosomes) was associated with increased risk of diabetes and partitioned the evidence for linkage to a greater degree than expected by chance, as assessed by simulation.

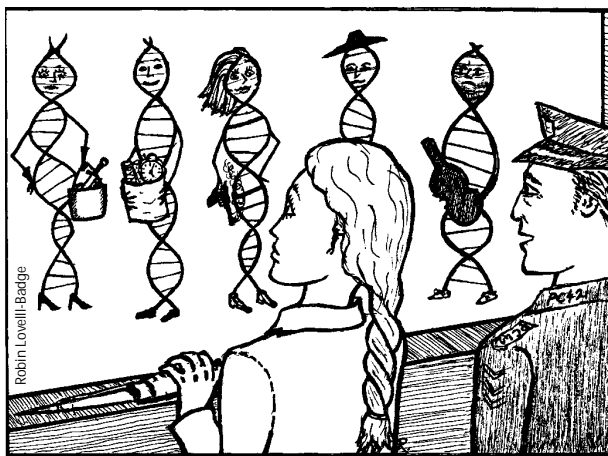
An unusual suspect

At first glance, UCSNP-43 seems an unlikely suspect for a polymorphism that affects disease causation (or even gene function). *CAPN10*—a ubiquitously expressed member of the calpain-like cysteine protease family—has no known or previously proposed role in diabetes physiology. Moreover, UCSNP-43 lies in an intron of *CAPN10*, and in a region that is not conserved in the mouse. It has no discernible effect on splicing. The authors are able to document an effect of UCSNP-43 on transcription, but the relevance of this effect to gene function or diabetes has yet to be established. Moreover, a polymorphism with a high frequency and modest association with increased risk (the G allele frequency is 75% in the general population and 80%

in diabetics) could not, acting alone, have produced the linkage signal observed in the original genome scan of 440 sibpairs. Indeed, the expected lod score attributable to such an allele is less than 0.05 and more than 100,000 sibpairs would be required to detect linkage with a lod score of 4.03, as observed by Hanis *et al.*³

These considerations raised the possibility that SNP43 might not be the sole perpetrator, but rather contributes modestly or is in LD with the actual causal variant(s). This led the authors to examine haplotypes in the region. Their data do not reveal a single variant or haplotype that is responsible for diabetes risk. Rather, heterozygosity for two different, common haplotypes—both of which carry the G allele at UCSNP-43—seems to be required to influence diabetes. Surprisingly, homozygotes for either haplotype were not at significantly altered risk. Critically, these findings were confirmed in an independent sample of Mexican Americans, making it unlikely that they represent a statistical fluctuation after a search over many haplotypes. It seems that the locus exists in several functionally distinct forms, and that a combination of two different forms in a patient may be required to influence disease.

Examination of genetic variation in *CAPN10* in two European populations offers a potential explanation why subsequent linkage studies (outside Mexican



Take your time, doctor. Which one do you reckon it is?

On the basis of distinct patterns of LD in the initial 10 individuals, 63 of the 179 polymorphisms were chosen for typing in a larger set of 100 diabetic cases and controls. Sixteen of these SNPs showed modest, nominally significant association with diabetes. To narrow the number of suspects, the authors developed a new statistical test, termed “partitioning of linkage.” The test assumes that the locus under study contributes to risk in only a subset of patients, and that this risk is conferred by a single common allele. The presence of this allele—or one in LD with it—should then selectively tag families responsible for evidence of linkage to the locus. Conversely, the remaining families (which lack the risk allele) should display no such evidence. In effect, the test demands association of an allele not just with disease but with those cases ‘linked’ to the locus under investigation. Presumably, associations that fail this test are spurious or too weak to be detected by linkage methods. Care must be taken to ensure that the test is applied in an unbiased fashion, and a comprehensive examination of its properties has yet to be published. Specifically, the appropriate statistical thresholds and the extent to which the method may be confounded by population admixture require investigation. In this case, however, the evidence for partitioning was strong, and implicated a single SNP (UCSNP-43), an intronic G/A polymor-

Americans) had not identified linkage to *NIDDM1*. In both German and Finnish patients, the same combination of two haplotypes shows a trend toward increased susceptibility, comparable in magnitude to that seen in Mexican Americans. However, although both haplotypes are common in Mexican Americans, one of them is rare in the Europeans, reducing the overall effect of the locus on risk of diabetes in these populations and potentially accounting for the weaker linkage findings. Examination of additional patient samples from these and other ethnic groups are critical to understanding the consistency and magnitude of these effects (and, based on the content of recent conferences, will probably be reported soon). In this regard, it is noteworthy that the first follow-up study⁵ fails to replicate the association with type 2 diabetes in Pima Indians (although it does report association with diabetic sub-phenotypes).

Learning from *NIDDM1*

Horikawa *et al.* have worked as long and as hard as any other group to clone a human polygene, and their report represents the state of the art of genetic analysis of complex traits. It is therefore both sobering and instructive that, despite their success in identifying associated haplotypes, answers to key questions remain uncertain. Does the association with *CAPN10* account for linkage to *NIDDM1*? Is *CAPN10* itself the 'diabetes' gene? If so, what specific polymorphism(s) is directly responsible for altering risk of diabetes, and by what mechanism?

The double-edged sword of LD

Initial localization to the *CAPN10* region relied on a scan of 21 SNPs across a 7-cM genetic interval. Is this density of markers adequate to comprehensively search for common haplotypes contributing to disease? At 1 SNP every 0.3 cM (equivalent to a genome-wide survey of 10,000 SNPs), this survey is significantly less dense than is expected to be necessary to search typical genomic regions for association⁶⁻⁸. As with the detection of linkage to Huntington disease in a screen of only 12 markers⁹, the positive finding of association with the sparse initial map must have been a pleasant surprise, and it is possible that other associations in the linked region have been overlooked. Just as linkage screens became more systematic with the development of genetic maps, it is hoped that dense SNP maps and efficient assays will similarly facilitate more comprehensive association screens. Determining the optimal density of markers, as well as appropriate thresh-

olds for following up initial associations, are important challenges for the field.

Of course, 1 SNP every 0.3 cM might suffice in cases where LD is extensive. However, in such situations the burden of a search is not lessened, only shifted. In cases where LD extends for tens to hundreds of kilobases, a sparse initial map may be sufficient to detect association (as it appears to have done in the present case); however, such an expanse of LD will require the exhaustive examination of many more candidate variants across this larger region. Many of these variants—potentially spanning several genes—may demonstrate similar or identical patterns of association with disease, making it difficult to determine which are causal variants and which happen to come along for the ride on the risk haplotype. For example, protection against HIV infection is conferred by inheriting a large, perfectly conserved, haplotype-block surrounding the gene *CCR5* (ref. 10). It is the biology of *CCR5* and its 32-bp deletion that proves which gene in the region grants protection.

Along these lines, it is intriguing that Horikawa *et al.* find 16 SNPs displaying some association to diabetes; these span much of the 66-kb sequenced region, and a number lie in *GPR35*, a gene adjacent to *CAPN10*. How many of these are markers for the described, associated haplotypes, and over what distance can the association be observed? If not attributable to the implicated haplotypes, what explains the association of the others? Additional work is required to understand the relationship between the three-marker system used initially to identify the sub-region (markers 1, 2, 19), the three markers for haplotype association in *CAPN10* (markers 43, 19, 63), and other positive markers throughout the region. For example, it is probably not a coincidence that marker 19 is present in both the initial 3-marker system and the haplotypes showing strongest association with disease. If so, could the at-risk haplotypes extend to UCSNP-1 and 2, which reside outside the 66 kb sequenced region?

In the end, genetic evidence alone may never be sufficient to establish the causal gene and mutation where LD is extensive. And even when the causal variant is known, regulatory effects (such as those proposed for UCSNP-43) can act over considerable distance. A regulatory element on human chromosome 5, for example, controls transcription of at least 3 genes spread over 120 kb (ref. 11). As Horikawa *et al.* point out, ultimate proof in such situations must be biological rather than statistical. Such proof can be obtained through the

transfer of specific variants, complete haplotypes or transcripts into cellular or animal models of disease. To observe an effect in such transgenic animals will require an appropriate host genotype and environment, however, which may be especially challenging in the case of *NIDDM1* and *CAPN10*. The current model requires compound heterozygosity for two different haplotypes at *CAPN10* and interaction with a locus on chromosome 15, a combination that will be hard to recapitulate in a heterologous system.

Closing arguments

The study by Horikawa *et al.* is a landmark in the decade-long effort to clone genes for polygenic disorders. It endorses the value of positional cloning for rounding up genes not otherwise suspected to have a role in a disease (rather than the usual suspects), and generates an exciting hypothesis for diabetes researchers to pursue. It also illustrates just how difficult the endgame of positional cloning can be, and contributes valuable data to the broader (often theoretical) debate about the utility of LD mapping and the analysis of common disease. (See, for example, the *Commentary*¹² on page 151 of this issue, and refs 13–15.) There is encouragement in Horikawa *et al.*'s successful application of SNPs and LD mapping to find common haplotypes displaying significant population-attributable risk. However, the complicated nature of genetic effects uncovered by the study indicates that we need to develop a better framework for connecting genotypes with phenotypes, and to contemplate genetic complexity as something more than 'several genes'. As other workers catch up with Horikawa and colleagues, the generality of these lessons will become clearer, and more surprises are sure to follow. As with any good detective story, the present instalment leaves one eager for the next. □

1. Kruglyak, L. & Lander, E.S. *Am. J. Hum. Genet.* **56**, 1212–1223 (1995).
2. Kruglyak, L. & Lander, E.S. *Am. J. Hum. Genet.* **58**, 1092–1093 (1996).
3. Hanis, C.L. *et al. Nature Genet.* **13**, 161–166 (1996).
4. Cox, N.J. *et al. Nature Genet.* **21**, 213–215 (1999).
5. Baier, L. *et al. J. Clin. Invest.* **106**, R69–R73 (2000).
6. Collins, A., Lonjou, C. & Morton, N.E. *Proc Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
7. Kruglyak, L. *Nature Genet.* **22**, 139–144 (1999).
8. Boehnke, M. *Nature Genet.* **25**, 246–247 (2000).
9. Gusella, J.F. *et al. Nature* **306**, 234–238 (1983).
10. Stephens, J.C. *et al. Am. J. Hum. Genet.* **62**, 1507–1515 (1998).
11. Loots, G.G. *et al. Science* **288**, 136–140 (2000).
12. Collins, F.S., Guyer, M.S. & Chakravarti, A. *Science* **278**, 1580–1581 (1997).
13. Weiss, K.M. & Terwilliger, J.D. *Nature Genet.* **26**, 151–157 (2000).
14. Lander, E.S. *Science* **274**, 536–539 (1996).
15. Risch, N. & Merikangas, K. *Science* **273**, 1516–1517 (1996).
16. Risch, N.J. *Nature* **405**, 847–856 (2000).