Direct isolation of polymorphic markers linked to a trait by genetically directed representational difference analysis

Nikolai A. Lisitsyn[1], Julia A. Segre[2], Kenro Kusumi[2], Natalia M. Lisitsyn[1], Joseph H. Nadeau[3], Wayne N. Frankel[3], Michael H. Wigler[1] & Eric S. Lander[2]

We describe a technique, genetically directed representational difference analysis (GDRDA), for specifically generating genetic markers linked to a trait of interest. GDRDA is applicable, in principle, to virtually any organism, because it requires neither prior knowledge of the chromosomal location of the gene controlling the trait nor the availability of a pre-existing genetic map. Based on a subtraction technique described recently called representational difference analysis, GDRDA uses the principles of transmission genetics to create appropriate Tester and Driver samples for subtraction. We demonstrate the usefullness of GDRDA by, for example, successfully targeting three polymorphisms to an interval of less than 1 cM of the mouse *nude* locus of chromosome 11.

[1]*Cold Spring Harbor Laboratory, P.O. Box 100, Cold Spring Harbor, New York 11724, USA*
[2]*Whitehead Institute for Biomedical Research, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA*
3 The Jackson Laboratory, Bar Harbor, Maine 04609, USA

Correspondance should be addressed to E.S.L.

Positional cloning, the isolation of genes based on their chromosomal location without prior knowledge of their biochemical function, is a powerful general approach that is applicable, in principle, to any organism[1]. Its actual use, however, has been much more restricted. Positional cloning depends on the ability to find tightly-linked genetic markers near a locus of interest, and hence the method has been practical only in the handful of organisms for which dense genetic maps have been constructed — principally, the fruit fly, nematode, mouse and human. For most organisms, genetic maps are either nonexistent or too rudimentary to allow routine positional cloning. To make positional cloning broadly applicable, one would ideally want a method for directly generating tightly-linked markers without recourse to a pre-existing genetic map. Here, we describe such a procedure, called genetically directed representational difference analysis (GDRDA).

Our method is based on a recently described subtractive technique called representational difference analysis (RDA) for identifying differences between two DNA samples, referred to as Tester and Driver[2]. Specifically, RDA is designed to clone restriction fragments that can be amplified by the polymerase chain reaction (PCR) from Tester but not Driver — either because the corresponding sequence is completely absent from the Driver due to a homozygous deletion or because it is contained in a small restriction fragment in the Tester but a large and, therefore, poorly amplifiable restriction fragment in the Driver. Thus, RDA can produce clones that detect restriction fragment length polymorphisms (RFLPs) between Tester and Driver.

To generate genetic markers linked to a trait, it is not enough simply to apply RDA to samples from a single affected and a single unaffected individual in a population or family. The abundant genetic variation among even close relatives in most populations, will mean that polymorphisms will likely be found throughout the genome. One requires a way to find polymorphisms specifically in the vicinity of the gene of interest. To ensure this, one needs Tester and Driver samples with the property that the Driver contains all of the alleles present in the Tester except in the region surrounding the target gene. As we describe below, such samples can be constructed by using classical transmission genetics. Although the methods are most easily applied to organisms that can be bred, they are applicable to natural populations as well.

Here, we describe two specific implementations of GDRDA. The first involves using congenic strains, while the second involves using progeny from an appropriate cross or pedigree. We tested the methods by using them to produce genetic markers linked to various mouse mutations and found them to be remarkably effective: of the one-third of clones that passed a simple initial screen, all (6/6) mapped to the desired region. Using congenic strains, genetic markers were produced near *pudgy* on chromosome 7 and *tottering* on chromosome 8. Using progeny from F_2 intercrosses, genetic markers were produced near *nude* on chromosome 11 and *staggerer* on chromosome 9. The GDRDA experiment with *nude* was aimed at finding polymorphisms within an interval of less than 1 cM around the locus. Three clones were produced and all mapped with 0.2 cM of *nude*, which comprised less than 1/2,000 of the mouse genome.

GDRDA with congenic strains

One ideal substrate for RDA would be a pair of congenic

strains[3] in which a particular gene has been transferred from one genetic background onto another by successive generations of backcrossing and selection. Congenic strains will be genetically identical except in a relatively small region surrounding the gene of interest. The region will typically be small enough to permit chromosomal walking to the target gene, but large enough to contain polymorphisms detectable by RDA. (RDA can detect only the minority of polymorphisms that cause gross differences in restriction fragments and thus, for example, comparison of isogenic strains that were identical except for a single mutation would likely fail to yield an RDA polymorphism.)

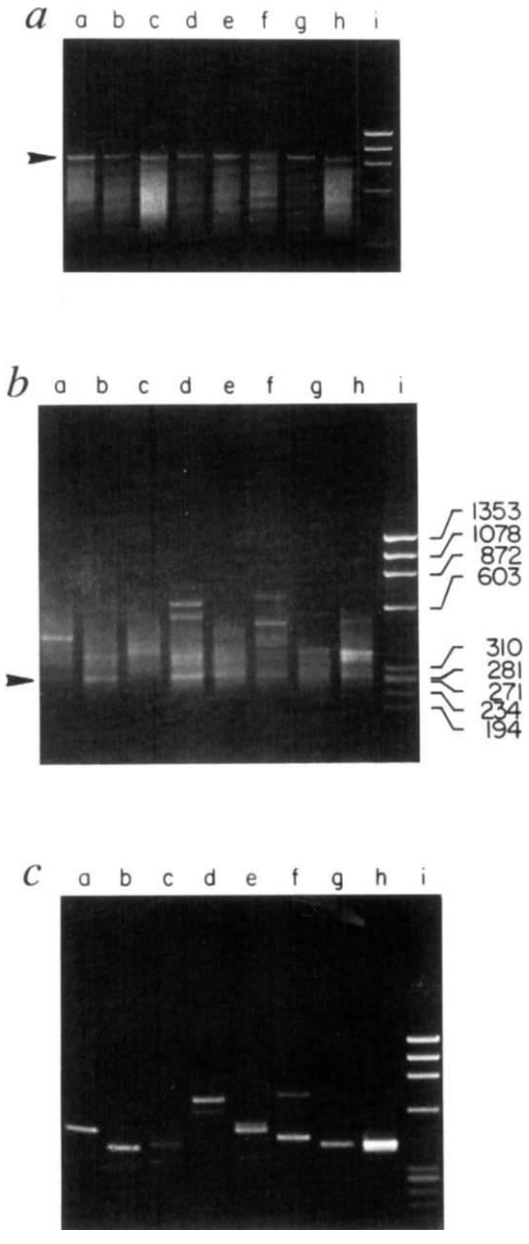

Fig. 1 Agarose gel electrophoresis of difference -products obtained after the first (*a*), second (*b*) and third (*c*) hybridization-extension-amplification steps in various experiments. Lanes are: *a*, *staggerer* cross; *b*, *Lurcher* congenic; *c*, *stargazer* congenic; *d*, *pudgy* congenic; *e*, *nude* congenic; *f*, *nude* cross; *g*, *severe combined immunodeficiency* congenic; *h*, *tottering* congenic; and *i*, *Hae*III digest of ϕX174 RF DNA. Sizes (bp) are indicated to the right. Arrows on the left show abundant mouse repeats removed by subsequent subtractions.

To test this implementation of GDRDA, we turned to the laboratory mouse, for which congenic strains have been developed for many interesting mutations. We selected congenic strains for *Lurcher* (*Lc*), *severe combined immunodeficiency* (*scid*), *pudgy* (*pu*), *tottering* (*tg*), *stargazer* (*stg*) and *nude* (*nu*). The congenic strains were constructed using between 11 and 40 generations of backcrosses (see Methodology for details of the strains).

RDA was performed in each case (see Methodology), using one of the pair of congenic strains as Tester and the other as Driver. Briefly, the first step involves preparing 'amplicons' from the Tester and Driver, which entails digesting each sample with a restriction enzyme, ligating the restriction fragments with a compatible adaptor, performing PCR using a primer complementary to the adaptor, and finally removing the adaptor by digestion with the original restriction enzyme. An amplicon contains only a portion of the genome, as it includes only small restriction fragments that are preferentially amplified. The Tester amplicons are then subjected to multiple rounds of hybridization-extension-amplification in the presence of excess Driver amplicon, under conditions favouring amplification of fragments present in the Tester amplicon that lack corresponding fragments in the Driver amplicon. Consequently, this procedure should yield small amplifiable restriction fragments which are present in Tester amplicons but absent or reduced in Driver amplicons. In these experiments, the restriction enzyme *Bgl*II was used and three cycles of hybridization-extension-amplification were performed. The resulting difference-products were separated by agarose gel electrophoresis. Several strong bands were visible upon staining with ethidium bromide, as well as a weak background smear (Fig. 1).

For each experiment, we cloned the difference product and selected six clones at random. We initially identified clones with distinct insert sizes (a total of 18 clones from the six experiments) and then characterized the clones by hybridizing them to Southern blots containing the Tester and Driver amplicons, to identify which clones showed the desired property of detecting a fragment in the Tester but not the Driver amplicon (Fig. 2). Of a total of 18 clones, this rapid test eliminated 15. The 'failures' could be grouped into three categories: First, seven clones detected a high-copy repeat in both Tester and Driver. Second, seven clones detected fragments in both the Tester and Driver amplicons. Finally, one clone failed to detect a signal in either Tester or Driver amplicon. Interestingly, all clones whose insert sizes did not correspond to one of the clear bands visible in the ethidium-stained difference product (11/18) failed the initial characterization. With a single exception, this was also true for the experiments described in the next section and suggests that this criterion might be useful for eliminating clones directly. Three clones (one each for *pudgy*, *tottering* and *stargazer*) showed the expected behaviour of hybridizing to the Tester but not the Driver amplicon. These three clones were then hybridized to Southern blots of Tester genomic DNA (as opposed to amplicon DNA) digested with *Bgl*II to determine whether they detected a unique genomic locus. Two clones (RDA-4.5 for *pudgy* and RDA 8.2 for *tottering*) detected a unique locus, whereas

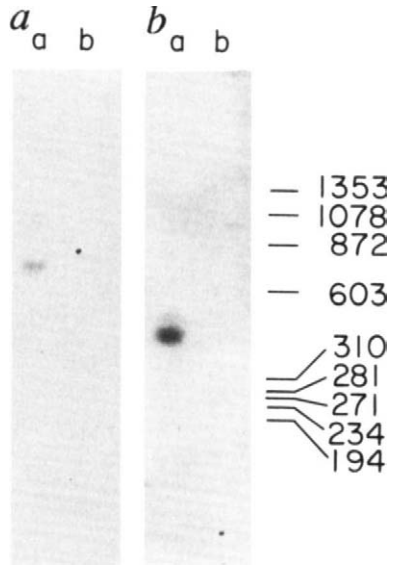

Fig. 2 Autoradiograms obtained after hybridization of probe RDA-4.5 (*a*) from *pudgy* congenic strains and probe RDA-8.2 (*b*) from *tottering* congenic strains to Southern blots containing *Bgl*II amplicons from Tester (lane a) and Driver (lane b). Sizes (bp) are indicated to the right. The faint band above the major DNA fragment is an unidentified PCR byproduct frequently observed on blots of *Bgl*II amplicons.

one clone (for *stargazer*) detected multiple loci and was eliminated. This rapid initial characterization thus eliminated all but two clones.

If GDRDA performed as intended, RDA-4.5 and RDA-8.2 should detect *Bgl*II polymorphisms mapping near *pudgy* and *tottering*, respectively. RDA-4.5 detected a *Bgl*II RFLP with a much smaller fragment in Tester than Driver (580 bp and 3.5 kb, respectively). Based on a genetic mapping panel consisting of 22 progeny from a (CAST/Ei × C57BL/6J-*mnd*)F2 intercross, this fragment mapped to the 9 cM interval between *D7Mit56* and *D7Mit25*, which is consistent with the location of *pudgy*[4]. Based on subsequent genetic mapping in a cross segregating *pu*, we determined that RDA-4.5 maps approximately 3 cM distal to *pu*, within the *pu-p* interval that was retained intact by the breeding scheme used to construct the stock (KK., W.F. and E.S.L., unpublished observations).RDA-8.2 detected a *Bgl*II RFLP with a much smaller fragment in Tester than Driver (400 bp and >3 kb, respectively). Using the same (CAST/Ei × C57BL/6J-*mnd*)F2 intercross as above, RDA-8.2 was found to map to the 7 cM interval between *D8MIT51* and *D8MIT9*, which is consistent with the location of *tottering*[4].

Thus, both GDRDA probes mapped to the desired region. Although the size of the target region differing between the congenic strains is not known precisely, it is estimated to be less than 15 cM based on the breeding schemes used in constructing the congenic strains. Accordingly, GDRDA successfully generated polymorphic probes in a region of less than 1% of the mouse genome around the target locus.

GDRDA with two-generation crosses

Congenic strains are an obvious choice for GDRDA, but they suffer from a major drawback. Producing congenic strains requires many generations of breeding, which can span years or decades depending on the organism. To develop a more practical and rapid approach, we devised a second implementation of GDRDA that requires only a simple two-generation cross.

Transmission genetics is used to produce a collection of

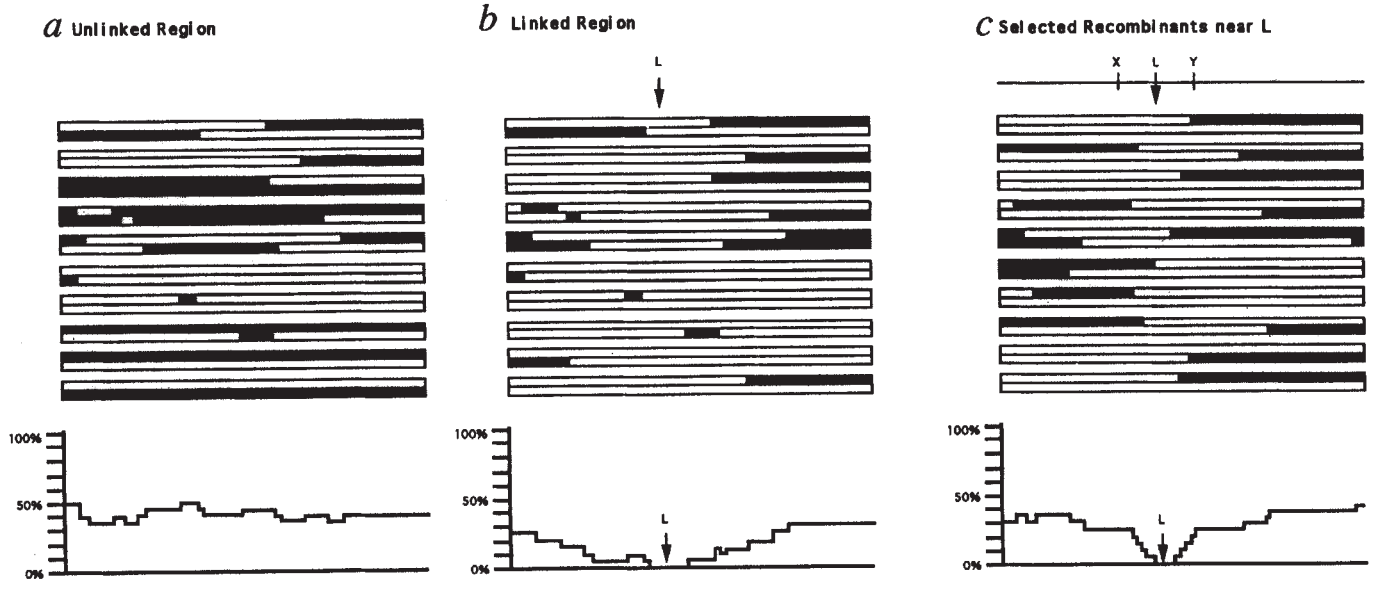

Fig. 3 Schematic diagram representing the principle underlying GDRDA with progeny from an F_2 intercross. Each panel shows hypothetical chromosomal genotypes from 10 progeny to be pooled to create a Driver; each chromosome is arbitrarily drawn to be 100 cM. Strain A carries a recessively-acting allele at locus L and is shown in white; strain B is shown in black. Graphs show percentage of B alleles present in Driver at each location along the chromosome. *a*, A chromosome unlinked to L (the percentage of B alleles remains close to 50%). *b*, The chromosome containing L, with progeny having the recessive phenotype selected at random (the percentage of B alleles dips slowly to 0% at L). *c*, The chromosome containing L, with progeny having the recessive phenotype selected to be recombinant between L and one of two flanking genetic markers, X or Y (the percentage of B alleles drops sharply to 0% in the X-Y interval).

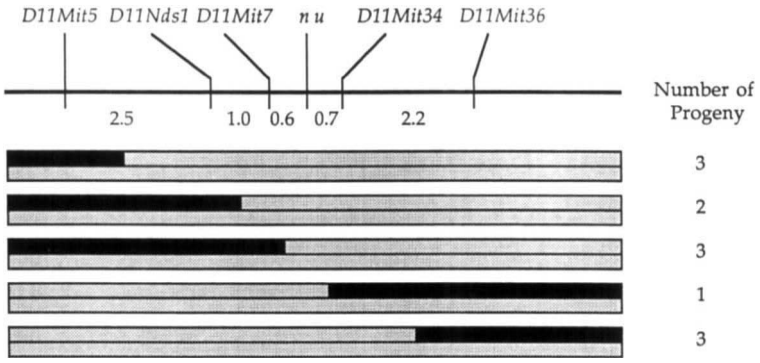

Fig. 4 Schematic diagram indicating chromosomal genotypes of the 12 (MOLF/Ei × AKR/J-*nu*^{str}) F2 progeny pooled to create the Driver for GDRDA, relative to a genetic map of polymorphic markers near the *nude* locus. Black indicates regions derived from MOLF/Ei. Shading indicates regions derived from AKR. The number of progeny of each type is indicated at the right.

siblings with the property that their pooled DNA is homozygous in the region of a target gene but heterozygous elsewhere in the genome. Let A and B denote two inbred strains differing at a target locus L of interest. (As discussed below, outbred strains can also be used with only minor modifications in the procedure.) Suppose that A carries a mutant allele *m* causing a recessive phenotype and B carries a wildtype allele + causing a dominant phenotype. For a Tester sample, one can use strain B itself. To create a Driver sample, one performs an F_2 intercross between the strains, selects a collection of k progeny showing the recessive phenotype, and mixes their DNA together. The principles of mendelian genetics predict that the Driver should contain: (i) no B alleles in the immediate vicinity of L, because progeny were selected for the recessive phenotype; (ii) a deficit of B alleles in a somewhat larger region around L, owing to linkage to L; and (iii) roughly equal proportions of A and B alleles elsewhere in the genome, because a collection of F_2 progeny should have genotypes AA, AB and BB in the ratio 1:2:1 at unselected loci (see Fig. 3*a*,*b*). If RDA is performed with this Tester and Driver, then one would expect that B alleles should be subtracted everywhere in the genome except in a region around L. GDRDA should thus yield polymorphic alleles from the wild-type chromosome at loci linked to L.

The targeting of the method can be somewhat improved in the event that the locus L has already been genetically mapped between two flanking genetic markers, X and Y (which might have been taken from a pre-existing genetic map or might have been generated by a previous application of GDRDA). For the Driver, one can select k/2 progeny in which a crossover had occurred between X and L and k/2 progeny in which a crossover had occurred between L and Y. This would guarantee that the proportion of B alleles is 25% at X and Y, ensuring that the region over which the proportion of B alleles is very low is restricted to the interval X–Y (Fig. 3*c*). As we demonstrate below, this refinement can allow targeting of very small intervals.

An important issue in the design of this experiment is the number of progeny that should be pooled. While the proportion of B alleles at unlinked loci in the Driver will have a mean value of 50%, the actual value will fluctuate across the genome. In the accompanying box, we discuss how many progeny should be pooled to ensure that the proportion of B alleles remains high enough throughout the genome to ensure efficient subtraction. For an F_2 intercross in the mouse, we conclude that 10 progeny should suffice.

To test this approach, we applied it to two mouse crosses involving the *nude (nu)* locus on chromosome 11 and the *staggerer (sg)* locus on chromosome 9. In both cases, GDRDA successfully generated probes mapping close to the target loci.

In the course of studies on *nude*, we had generated 416 (MOLF/Ei × AKR/J-*nu*^{str}) F_2 intercross progeny, genotyped them for various genetic markers on Chromosome 11 and determined the position of *nude* relative to these markers (J. S., J.N., Benjamin Taylor and E.S.L., unpublished data). Using this information, we selected 12 nude progeny having crossovers between *nude* and closely linked markers (Fig. 4). All of the crossovers occurred within a 7 cM interval defined by *D11Mit5* and *D11Mit36*, and 4 of the 12 occurred within a 1.3 cM interval defined by *D11Mit7* and *D11Mit34*. A Driver sample was prepared by pooling equal amounts of DNA from these 12 progeny; the corresponding Tester sample was DNA from the MOLF/Ei parental strain. In principle, GDRDA should produce MOLF/Ei alleles of polymorphisms in the interval *D11Mit5* and *D11Mit36*. Moreover, if the proportion of B alleles outside this interval sufficed to allow efficient subtraction, the polymorphisms might be targeted preferentially to the small interval between *D11Mit7* and *D11Mit34*.

Using this Tester and Driver combination, we performed RDA with the restriction enzyme *Bgl*II. In the resulting difference product, two clear bands (700 bp and 450 bp) were visible by ethidium bromide staining. These were cloned to produce probes RDA-6.1 and RDA-6.2. As above, the probes were initially characterized by hybridization to Southern blots of Tester and Driver amplicons. RDA-6.1 turned out to detect a large number of bands in both amplicons and was eliminated. RDA-6.2 showed the expected pattern of hybridizing to the Tester but not Driver amplicon. The probe was then hybridized to Southern blots of mouse DNAs digested with *Bgl*II. It detected an RFLP with a 450 bp allele in MOLF/Ei and a 4 kb allele in AKR/J-*nu*^{str}. Using this RFLP, the locus detected by RDA-6.2 was genetically mapped. To obtain approximate localization, we genotyped 20 (MOLF/Ei × AKR/J-*nu*^{str}) F_2 progeny that showed no recombination between genetic markers flanking *nude* and found that the RFLP showed an inheritance pattern completely concordant with that of the *nude* locus itself (Fig. 4). To obtain finer localization, we then genotyped the 12 nude F_2 progeny used to create the Driver and found that the RFLP again showed complete concordance with *nude*—i.e., the progeny were all homozygous for the AKR allele of the RFLP. This proves that RDA-6.2 maps within the 1.3 cM interval bounded by *D11Mit7* and *D11Mit34*. Subsequent analysis of additional F_2 progeny (J.A.S., J.H.N., Benjamin Taylor and E.S.L., unpublished data) has shown that RDA-6.2 recombined with *nude* only twice in 1290 meioses, corresponding to a genetic distance of only 0.2 cM. Thus, GDRDA successfully targeted a probe to a region less than 1/2,000 of the mouse genome.

We next attempted to generate additional clones by repeating GDRDA using the restriction enzyme *Bam*HI. Two of three clones, RDA-10.2 and RDA-10.4, showed

the expected pattern of hybridizing to the Tester but not the Driver amplicon. Both probes detected RFLPs between MOLF/Ei and AKR/J-*nu*^{str} (with allele sizes of 600 bp and 4–5 kb for RDA-10.2 and 500 bp and 3 kb for RDA-10.4). Genetic mapping subsequently showed that both probes mapped close to *nude*. The 12 F_2 progeny used to create the Driver were all homozygous for the AKR allele for both RFLPs, indicating that both loci mapped in the 1.3 cM interval between *D11Mit7* and *D11Mit34*. Subsequent analysis of additional F_2 progeny (J.A.S., J.H.N., Benjamin

Box1 Experimental design issues

In applying GDRDA to an F_2 intercross, how many progeny should be pooled to create the Driver? The method requires that the proportion of B alleles is sufficient to ensure subtraction at all unlinked loci. While the expected proportion will be 50% by mendelian segregation, the actual proportion will fluctuate across the genome. The more progeny pooled, the smaller the fluctuations.

The critical proportion α of B alleles needed to ensure subtraction at a locus is not known precisely and can only be determined based on empirical evidence from many RDA experiments. Indeed, it may depend on the nature of the sequence, the hybridization conditions used and the ratio of Tester and Driver at each stage. Nonetheless, the current RDA protocol — which employs a 80-fold excess of Driver on the first round — seems to allow efficient subtraction of alleles present at 10–15% in the Driver (N.L., unpublished data). Thus, one might set the critical threshold for subtraction at $\alpha = 0.10$–0.15.

Given a choice for critical threshold α, how many progeny k should be pooled? Let $a_n(\alpha)$ denote the expected length of the region linked to L for which the proportion π_B of B alleles is less than α and $b_n(\alpha)$ denote the expected length of *unlinked* regions of the genome for which $\pi_B < \alpha$, where n is the number of recombinant haploid genomes pooled to create the Driver (that is, $n = 2k$). The ratio $c_n(\alpha) = \alpha_n(\alpha)/b_n(\alpha)$ should give a good indication of the ratio of linked to unlinked clones that should be produced by RDA. (For the calculation of $c_n(\alpha)$, see Methodology.) If $c_n(\alpha) >> 1$, then linked clones should consititute the majority of fragments surviving subtraction. If $c_n(\alpha) < 1$, linked clones will be a minority which must be identified by subsequent screening. The number of progeny should thus be $k \geq n/2$, where n is the smallest integer such that $c_n(\alpha) \geq C$, for a chosen lower bound C.

A graph of $c_n(\alpha)$ is shown in Fig. 5. Applying this to the mouse genome (genetic length ≈ 16 Morgans) and choosing a critical threshold $\alpha = 0.15$ for subtraction, one has $c_n(0.15) = 7.7, 13.6, 24.1$, and 182.6 for $k = 6, 8, 10$ and 12 F_2 progeny pooled (with $n = 2k$). To ensure $c_n(\alpha) >> 1$, it might thus be prudent to pool at least 10 F_2 progeny.

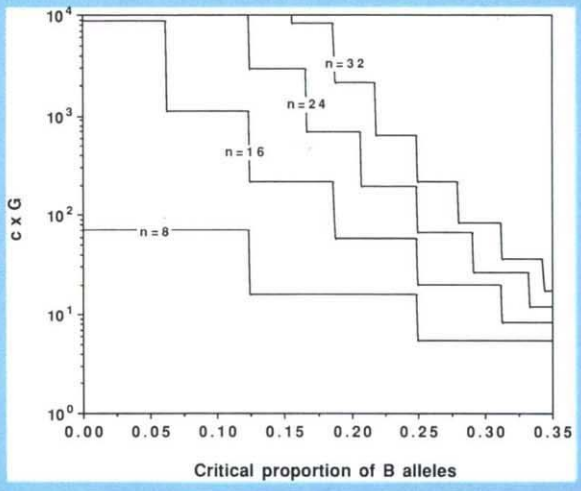

Fig. 5 Mathematical analysis of construction of Driver for an F2 intercross. Curves show the value of $c_n(\alpha) \times G$, where G is the length of the target genome measured in Morgans and $c_n(\alpha)$ is closely related to the ratio of linked to unlinked probes expected to occur in the difference-product, assuming a critical proportion α of B alleles necessary for efficient subtraction and a Driver pool containing n haploid genomes (see text for exact description). Given α, the number n should be chosen to ensure that $c_n(\alpha) >> 1$ or, using the graph, that $c_n(\alpha) \times G >> G$. (Note, the function $c_n(\alpha) \times G$ is given because, unlike $c_n(\alpha)$ itself, the value does not depend on G.) The curves are step functions because, if n haploid genomes are pooled, the proportion of B alleles at any locus must be an integral multiple of $1/n$.

How close to L will the linked polymorphisms be? Assuming the simple model of a critical threshold α for subtraction, they would be expected to lie roughly within recombination fraction $\theta \approx \alpha$ of L. If $\alpha = 0.10$, the target interval thus may be about 20 cM. If subtraction is not all-or-none, there should be a bias toward the centre of the region because the proportion of B alleles will be lowest there. If progeny can be selected having recombinations near L (as discussed in the text), the interval targeted can be made much smaller.

GDRDA can be applied to a backcross between inbred strains with only a minor modification in the analysis. In a (A x B)F1 x A backcross, use the (A x B)F1 animals as Tester and use a collection of k backcross progeny showing the recessive phenotype as the Driver. The Tester and Driver for these k backcross progeny are identical to those that would be obtained by taking the Tester and Driver for k/2 F_2 intercross progeny and mixing each 1:1 with strain A; the mixing with strain A should have no effect, since A alleles should be efficiently subtracted. The number of progeny to pool is thus twice as many as that for the corresponding intercross (that is, $k = n$, where n is the smallest integer such that $c_n(\alpha) \geq C$).

GDRDA can also be applied to crosses involving non-inbred matings. Consider a mating in an outbred population between individual C who is heterozygous ($m/+$) and another individual D. The Tester can be C and the Driver can be a collection of progeny who inherited the allele m from C. If m causes a dominant phenotype, these progeny can be readily identified based on their phenotype. If m causes a recessive phenotype, they could identified either by progeny testing or by using a parent D who is also heterozygous and selecting homozygous progeny. Subtraction should yield alleles present only on the chromosome carrying the + allele in C. The situation differs from a backcross with inbred strains only in one respect: one must ensure subtraction of two possibly different alleles at unlinked loci in C. To account for regions in which the proportion of either allele is too low, the function $b_n(\alpha)$ should be replaced by $2b_n(\alpha)$. The minimum number of backcross progeny that should be pooled is thus $k = n$, where n is the smallest integer such that $c_n(\alpha)/2 \geq C$. This is only a slight increase over the corresponding backcross.

Finally, the progeny in the outbred matings need not be full sibs. One could use progeny from matings of C to multiple partners, D_1, D_2, ..., D_j. The potential drawback is that a linked C allele could be subtracted if it is present in any of the D_j, which would decrease the number of detectable polymorphisms as j increases. The half-sib design may be especially convenient in the case of livestock, for which a single male is often mated to multiple females. □

Taylor and E.S.L., unpublished data) showed that RDA-10.4 recombined with *nude* only once and that RDA-10.2 never recombined with *nude* in 1290 meioses. In summary, GDRDA produced three distinct polymorphisms mapping within 0.2 cM of the target locus.

Finally, we performed an analogous experiment using the *staggerer* mutation. In another project, we had genotyped 270 (C57BL/6J-*sg* × DBA/2J)F2 intercross progeny for various genetic markers on chromosome 9 and determined the position of *staggerer* relative to these markers (K.K., W.N.F., Muriel Davisson, and E.S.L., unpublished data). Using this information, we selected 13 *sg/sg* homozygous progeny having crossovers in a 10 cM interval containing *staggerer* defined by *D9Mit48* and *D9Mit9*. A Driver sample was prepared by pooling DNA from these 13 progeny; the corresponding Tester sample was DNA from the DBA/2J parental strain. In principle, GDRDA should produce DBA/2J alleles of polymorphisms in the interval between *D9Mit49* and *D9Mit11*.

RDA was performed with the restriction enzyme *Bgl*II. A single strong band (500 bp) was visible by ethidium bromide staining and was cloned to produce the probe RDA-1.1, which passed the initial characterization tests. The probe detected a *Bgl*II RFLP between Tester and Driver (500 bp allele in DBA/2J and >4 kb in C57BL/6J-*sg*). The RFLP mapped to the interval between *D9Mit49* and *D9Mit11* based on a (CAST/Ei × C57BL/6J)F2 intercross. When the 12 recombinant progeny that had been used in the Driver were genotyped, we found that 9 progeny were homozygous for the C57BL/6J-*sg* allele but three progeny were heterozygous. In contrast to the *nude* experiments in which all three probes derived from a region for which the Driver completely lacked Tester alleles, this experiment yielded a probe from a region near *sg* for which the Driver contained the Tester allele at a proportion of 11.5% (that is, 3/26). Subsequent genotyping of the 270 (C57BL/6J-*sg* × DBA/2J) F2 progeny has shown that RDA-1.1 maps approximately 4.5 cM distal to *sg* (K.K., W.N.F., Muriel Davisson and E.S.L., unpublished data). In summary, GDRDA produced closely linked markers in both the *nude* and *staggerer* crosses.

Discussion

GDRDA is unique among molecular genetic techniques in that it provides a way to target DNA probes to the vicinity of a gene without prior knowledge of either the gene's function or position. By applying classical transmission genetics, one can prepare DNA samples from mixtures of progeny that differ only near the gene of interest and then use the powerful subtraction technique of RDA to clone these differences. The technique opens the prospect of genetic analysis and positional cloning even in organisms without pre-existing genetic maps.

We describe two particular implementations of GDRDA, using congenic strains and two-generation crosses. Both approaches successfully produced probes mapping near various target genes. Indeed, every clone (6/6) that passed a rapid initial characterization (detecting a unique fragment in Tester but not Driver amplicon and a unique locus in genomic DNA) mapped to the desired location. In the case of the *nude* cross, we obtained three different probes that mapped within 0.2 cM of the target locus.

The yield of probes was relatively low (6 probes from 9 experiments), which is perhaps not surprising in view of the multiple rounds of exponential competition among PCR products during RDA. The number of probes might be increased through the use of additional restriction enzymes for amplicon preparation, as demonstrated by the successful use of *Bam*HI in the case of the *nude* experiment. Some restriction enzymes, such as *Taq*I, may produce a higher yield of polymorphisms. It may also be possible to generate new clones with a single restriction enzyme by blocking the amplification of already-identified clones by adding them back to the Driver. Finally, it may be possible to detect less drastic changes in the length of restriction fragments by initially fractionating Tester and Driver by gel electrophoresis and performing subtraction on specific size fractions.

Application of GDRDA to congenic strains is straightforward. However, the real power of GDRDA lies in its application to crosses, because the breeding or pedigree collection required is within the realm of practicality for a wide range of organisms. The technique can be applied to any trait whose presence implies homozygosity for a particular allele at a trait-causing locus, so that these homozygotes may be pooled to create a Driver.

An interesting feature of the application to crosses is that the targeting of GDRDA can be improved by successive iteration. Given a large cross, one could first generate flanking markers that are linked, but perhaps not very closely, to the target locus L. Using such flanking markers to identify recombinant progeny, one could perform subsequent subtractions with these progeny to target successively smaller intervals. As shown in the case of the *nude* and *staggerer* crosses, the use of recombinant progeny can effectively target quite small intervals. The ultimate resolution of this approach should be limited only by the actual density of polymorphisms detectable by GDRDA; we estimate this density to be 1–2 per megabase for an enzyme such as *Bgl*II.

We have focussed here on the application of GDRDA to F_2 intercrosses between inbred strains, but the technique is more broadly applicable. It can be applied to backcrosses between inbred strains, two-generation families in an outbred population (for organisms for which inbred lines are not available), and half-sib mating schemes (common in livestock breeding). Considerations in designing such experiments are discussed in the accompanying box.

The application of GDRDA to random-breeding populations should include the analysis of human families. One might, for example, use an individual affected with a dominant disease as Tester and a collection of unaffected close relatives as Driver. In some families, there may be too few relatives to ensure subtraction of all unlinked regions. In such cases, GDRDA should at least enrich for linked probes which could then be subsequently screened for linkage. We will discuss this issue in more detail elsewhere.

Notwithstanding continuing advances in genomic analysis[5], construction and application of dense genetic linkage maps remains a daunting task. GDRDA offers the prospect of obviating the need for such maps, at least for certain purposes. In particular, GDRDA should open the prospect of genetic mapping and positional cloning of monogenic traits in most experimentally and agriculturally important animals, plants and fungi.

Methodology

Mouse strains. All mouse strains used were maintained at The Jackson Laboratory, with the exception of those used for the *Lc* congenic experiment, which were maintained at the Wadsworth Center, Albany, NY and provided by Anne Messer. The congenic

strains used were: *Lurcher* (*Lc*, chromosome 6). This dominantly-acting mutation arose in the Mi^{wh} stock and a congenic strain was produced by 40 generations of backcrosses to BALB/cBy. The Tester was a *Lc*/+ female from the BALB/cBy congenic strain (N40) and the Driver was a BALB/cBy female. *Severe combined immunodeficiency* (*scid*, chromosome 16). This recessively-acting mutation arose on C.B-17 (a BALB/c-like strain) and a congenic strain was produced by 11 generations of backcrosses to C3H/HeJ. The Tester was a C3H/HeJ-*scid* male (N11) and the Driver was a C3H/HeJ male. *Pudgy* (*pu*, chromosome 7). This recessively-acting mutation arose on a non-inbred stock. It was maintained on a homozygous chincilla (c^{ch}) stock, in *trans* to the nearby *p* mutation (that is, $pu + c^{ch}/+ p\ c^{ch}$) and was subsequently brother-sister mated for 42 generations with selection for heterozygotes in every other generation. The breeding scheme should maintain two alternative forms of a congenic region including the *pu-p* interval, but the animals should be identical outside this region. The Tester was a *pu*/+ female and the Driver a *pu*/*pu* female from this stock (N42). *Tottering* (*tg*, chromosome 8). This recessively-acting mutation arose on a DBA/2J genetic background and a congenic strain was produced by 34 generations of backcrosses to C57BL/6J. The Tester was C57BL/6J-*tg* female (N34) and the Driver was a C57BL/6J female. *Stargazer* (*stg*, chromosome 15). This recessively-acting mutation arose on a A/J background and a congenic strain was produced by 19 generations of backcrosses to a (C3H/HeJ × C57BL/6J) hybrid background. The Tester was a *stg/stg* female from the congenic strain (N19) and the Driver was a 1:1 mixture of C3H/HeJ and C57BL/6J female DNA. *Nude* (chromosome 11). This recessively-acting mutation arose in a non-inbred strain and a congenic strain was produced by 12 generations of backcrosses to C57BL/6J. The Tester was a C57BL/6J female and the Driver was a C57BL/6J-*nu* female (N12). For further information about the mutations discussed in this paper, see ref. 6.

RDA procedure. RDA was performed essentially as described[2]. A detailed protocol is available directly from the authors. To maximize the success of RDA, it may be helpful to employ the following controls: (i) ligation of PCR products with new adaptors on each round of RDA can be monitored by PCR and subsequent gel electrophoresis before hybridization, which should show a detectable increase in fragment size distribution; (ii) concentration of Tester and Driver DNA at each step should be determined by gel electrophoresis, using *Sau*3A digested human DNA as a control; (iii) experiment 1 from ref. 2 can be performed in parallel with the main experiment, as a positive control.

In this work, all amplicons were prepared by digesting 2 μg each of Tester and Driver DNAs with either *Bgl*II or *Bam*HI. The iterative hybridization-extension-amplification step was repeated three times. The resulting material was digested with the same restriction enzyme as used to prepare the amplicon, ligated to *Bam*HI-digested and dephosphorylated pBluescript II SK(−), and transformed into *E. coli* XL-Blue competent cells according to the supplier's recommendations.

Initial characterization of RDA clones. For each experiment, six white colonies were picked at random and the inserts were immediately analyzed by PCR. The colonies were resuspended in 100 μl LB medium containing ampicillin (for subsequent growth and plasmid isolation) and a 5 μl aliquot was immediately transferred to 100 μl of a PCR reaction containing 1 μM each of Seq24 primer (5'-CGACGTTGTAAAACGACGGCCAGT-3') and Rev25 primer (5'-CACACAGGAAACAGCTATGACCATG-3'), 67 mM Tris-HCl (pH 8.8 at 25 °C), 4 mM $MgCl_2$, 16 mM $(NH_4)_2SO_4$, 10mM β-mercaptoethanol, 170 μg ml^{-1} bovine serum albumin and 200 μM (each) of dATP, dGTP, dCTP and dTTP. The mixtures were incubated at 95 °C for 5 min and cooled to 72 °C, after which 5 U of Ampli*Taq* polymerase (Perkin-Elmer Cetus) are added and the mixture was thermocycled for 30 cycles (95 °C for 1 min, 72 °C for 3 min) followed by a final incubation at 72 °C for 10 min. The amplified plasmid inserts were analysed by agarose gel electrophoresis to identify those having distinct sizes. These were purified on Qiagen-tip20 columns (Qiagen Inc.), according to supplier's recommendations. To determine whether the clones represented sequences which were selectively present in the Tester but not Driver amplicons, selected inserts were radioactively labelled using a Megaprime DNA labelling system (Amersham) according to the supplier's recommendations, and hybridized to Southern blots containing DNA from Tester and Driver amplicons, which had been electrophoresed in a 2% agarose gel and transferred using a vacuum blotting apparatus to GeneScreen Plus membranes. Finally, clones were tested to determine whether they detected a unique genomic locus by hybridizing them to Southern blots of restriction-digested genomic DNA, with washing at moderate stringency (two 30 min. washes in 0.5× SSC, 0.1% SDS at 65 °C).

Genetic mapping of RFLPs. Clones detecting a fragment present in Tester but not Driver amplicons were hybridized to Southern blots containing restiction-digested mouse genomic DNA to test whether they detected a RFLP between Tester and Driver. Clones detecting RFLPs were subsequently genetically mapped in the mouse genome, by hybridizing them to Southern blots containing restriction-digested DNA from progeny of various two-generation mouse crosses. Southern blotting and hybridization was essentially as described[7]. The inheritance pattern of the RFLPs was compared to that of various simple sequence length polymorphisms (SSLPs) that mapped to the regions of interest. The SSLPs and the genotyping protocol were previously described[8,9].

Calculation of $c_n(\alpha)$. As described in Box 1, the proportion of linked to unlinked clones produced by GDRDA depends on the ratio $c_n(\alpha) = a_n(\alpha)/b_n(\alpha)$, where $a_n(\alpha)$ is the expected length of the region linked to L for which the proportion π_B of B alleles is less than α, $b_n(\alpha)$ is the expected length of the unlinked regions of the genome for which $\pi_B < \alpha$, and n is the number of recombinant haploid genomes pooled (that is, $n = 2k$, where k is the number of F2 progeny pooled). The function was calculated as follows:

$$a_n(\alpha) = \int_{-\Delta}^{\Delta} p_n(\alpha,\theta(x))\,dx \quad \text{and} \quad b_n(\alpha) = p_n(\alpha,0.5)\,G,$$

where

$$p_n(\alpha,x) = \sum_{i=0}^{i\le\alpha n} \binom{n}{i} x^i (1-x)^{n-i}$$

denotes the cumulative probability distribution of the binomial distribution for n independent events having probability x, θ(x) is a chosen map function converting genetic distance to recombination frequency (we used Haldane's map function, $\theta(x) = (1 - e^{-2x})/2$), Δ denotes the maximum distance that should be considered to be 'linked' to L (we used Δ = 0.50 Morgans), G denotes the genetic length of the 'unlinked' genome, and all lengths are measured in Morgans. The first equation follows immediately from the definition of $p_n(\alpha,x)$. The second equation follows by first noting that the proportion of B alleles at distance x from L is binomially distributed with probability equal to the recombination fraction θ(x) and then integrating over the 'linked' points in an interval [-Δ, Δ] centred on L.

Acknowledgements

We thank L. Rodgers, M. Riggs, J. Smith, A. Weaver, D. Varnum, J. Duffy, M. Okler, A. Messer, L. Shultz and M. Davisson for their help and contribution to this work. W.N.F. was a Special Fellow of the Leukemia Society of America. This work was supported in part by grants from the National Center for Human Genome Research, the National Science Foundation, and the Markey Foundation to E.S.L., the National Institute of Child Health and Development to J.H.N. and the National Cancer Institute and the American Cancer Society to M.H.W.

1. Collins, F. Positional cloning: let's not call it reverse anymore. *Nature Genet.* **1**, 3–6 (1992).
2. Lisitsyn, N., Lisitsyn, N. & Wigler, M. Cloning the difference between two complex genomes. *Science* **259**, 946–951 (1993).
3. Snell, G.D. Histocompatibility genes of the mouse. II. Production and analysis of isogenic resistant lines. *J. natn. Cancer Inst.* **21**, 843–877 (1958).
4. Hillyard, A.L., Doolittle, D.P., Davisson, M.T., Maltais, L., & Roderick, T.H. Locus map of the mouse with comparative map points of human on mouse. (GBASE Electronic Database, Jackson Laboratory, Bar Harbor, Maine, March, 1993).
5. Williams, J.G.K., Reiter, R.S., Young, R.M. & Scolnick, P.A. Genetic Mapping of mutations using phenotypic pools and mapped RAPD markers. *Nucl. Acids Res.* **21**, 2697–2702 (1993).
6. Lyon, M.F. & Searle, A.G. *Genetic Variants and Strains of the Laboratory Mouse* 2nd edn (Oxford Univeristy Press, Oxford, 1989).
7. Brown, T. in *Current Protocols in Molecular Biology* (eds Ausubel, F. *et al.*) 2.9.1–2.9.15 (Green, New York, 1993).
8. Dietrich, W. *et al.* A Genetic Map of the Mouse Suitable for Typing Intraspecific Crosses. *Genetics* **131**, 423–447 (1992).
9. Whitehead Institute/MIT Center for Genome Research, Mouse Genome Public Electronic Database, "genome_database@genome.wi.mit.edu" (July, 1993).

Received 17 August; accepted 18 October 1993.