RESTRICTION FRAGMENT LENGTH POLYMORPHISMS AND GENETIC IMPROVEMENT

Polimorfismo de la restricción por longitud de fragmentos y mejora genética

M. SOLLER*
J.S. BECKMAN**

ISRAEL

Methods have recently been developed, based on the use of restriction enzymes and cloned DNA, that allow a new class of genetic polymorphisms: restriction fragment length polymorphisms (henceforth, RFLPs) to be detected (Botstein et al., 1980; Solomon and Bodmer, 1979). Large number of RFLPs have already been demonstrated in humans (Bell et al., 1981; Jeffreys, 1979; Wyman and White, 1980) and in other organisms (Petes and Botstein, 1977).

The purpose of this paper is to draw attention to the potential usefulness of this class of polymorphisms in genetic improvement. We will describe (1) the nature of RFLPs and the methodologies involved in their detection, (2) the implications of a virtually unlimited source of genetic polymorphisms for breeding practice, in particular for improvement of strain crosses, and for within population selection, and (3) present costs of RFLP methodologies and anticipated costs of their applications.

I. RFLP METHODOLOGIES

Restriction fragments

Restriction endonucleases are enzymes that recognize specific nucleic acid sequences in DNA and cleave the DNA at these sites or at adjacent sites (Zabeau and Roberts, 1979). Depending on their specificity, most restriction endonucleases will have 4 or 6-base recognition sequences. Such sequences can be expected to occur once in every $4^n$ or $4^6$ nucleotides, respectively.

The more prevalent the recognition sequence is in the DNA, the more frequently it will be cleaved by the enzyme which recognizes that sequence. The DNA fragments so formed, can be separated by gel electrophoreses since smaller fragments migrate more rapidly through the gel than larger fragments. When the DNA from a higher organism is digested by restriction enzymes, so many different sized fragments are produced that a continuous smear is formed on the gel. Specific fragments are therefore detected by the use of an appropriate probe. This is carried out as follows: The DNA pattern is transferred from the gel to a solid support, such as a nitrocellulose filter (Southern, 1975). The DNA, now bound to the filter is then exposed to a radioactively labelled probe under conditions that promote DNA-DNA hybridization. The probe used would be a cloned DNA sequence homologous to a particular DNA fragment or some portion of it. The unhybridized radioactivity is then washed away and the filter is dried and placed against photographic film for autoradiographic exposure. After film development the specific fragment that hybridized with the probe will be visualized as a band on the film.

*Department of Genetics, The Hebrew University, 91904, Jerusalem, Israel.
**Department of Field and Garden Crops, Agricultural Research Organization, 50200 Bet Dagan, Israel.
By using specific probes (say reverse transcribed unique DNA) a DNA fragment that occurs as rarely as once in a million or less can be detected. The probes used in RFLP detection do not have to be homologous with known genes. Any unique sequence will suffice as long as it hybridizes with some part of one of the DNA fragments formed by the restriction process.

**Restriction fragment length polymorphisms**

Base changes can alter the sequences that are recognized by restriction enzymes, abolishing sites or creating new sites for particular enzymes. Deletions or transpositions of large elements into a site will make simultaneous changes in the restriction patterns of a number of enzymes. As a result, a given restriction enzyme will not always cleave a given DNA molecule at the same point in two individuals. In this case fragments of different length will be formed when the DNA of the two individuals is digested. The unequal sized fragments will travel at different rates through the gel, and following hybridization and autoradiography, the bands formed will be located at different points on the film. In this manner a restriction fragment length polymorphism will have been demonstrated. A hypothetical example is illustrated in Figure 1.

**Figure 1.** Hypothetical example illustrating a restriction fragment length polymorphism. A region of homologous chromosomal DNA sequence from two individuals is diagrammed. Both segments contain a sequence (indicated by the box with hatched lines) that has been cloned. The site for Enzyme R on the right side of the box in individual 2 has been abolished by alteration of the DNA sequence: A larger fragment will therefore be produced when DNA from individual 2 is digested with enzyme R. Autoradiographic visualization of the fragments containing the cloned sequences are shown in the lower half of the figure.

**Restriction Sites in Chromosomal DNA**

1)  

2)  

Restriction fragments visualized after hybridization with a cloned probe

Origin

1 2

(-)

(+)

397
Expected frequency and mode of inheritance of RFLPs

A large variety of restriction enzymes having different specificities are available and RFLPs in both coding and non-coding DNA sequences can be detected with appropriate probes. Thus the RFLP method can potentially monitor a significant fraction of the total genome and should be virtually unlimited in its ability to detect large numbers of polymorphisms.

Careful consideration of Figure 1 shows that in a heterozygote for both types of DNA molecule, both bands would be present in the autoradiograph. Thus RFLPs can be expected to show a codominant mode of inheritance, and this has indeed been found in a number of instances (Petes and Botstein, 1977; Wyman and White, 1980). Furthermore, since RFLPs will often be located in flanking sequences or in introns, one would expect them to lack secondary pleiotropic effects on quantitative characters. This class of polymorphism should therefore be well suited for purposes of genetic analysis and improvement of quantitative traits.

The number of RFLPs required for complete genome coverage

A minimum estimate of the number of properly placed markers for complete genome coverage would be T/c, where T is the total map length in cM, and c is the spacing between markers in cM. For purposes of genetic analysis of quantitative traits, a spacing of 40 cM between markers would be sufficient (Soller et al., 1976). For purposes of marker assisted introgression or selection, a spacing of 20 cM between markers would be desirable (Soller and Plotkin, 1978).

The total number of polymorphic markers that need to be detected in order to obtain a well spaced subset has been considered in detail for the human genome by Bishop et al. (to be published). From their tables it appears that roughly twice the minimum number of polymorphisms will give a subset of markers that will cover 90-95% of the genome at the desired spacing. On this basis for dairy cattle say (T=3000), 75 well placed markers would be required for a 40 cM spacing and 150 for a 20 cM spacing, while the total number of polymorphisms that would have to be detected to yield an appropriately spaced subset would be about 150 to 300, respectively.

II. RFLPs AND GENETIC IMPROVEMENT

In this section we will describe two possible applications of RFLPs to genetic improvement in animal breeding. The first, will deal with improvement of existing strain crosses or of crosses between inbred lines, as in poultry. The second will deal with increased effectiveness of young bull selection in dairy cattle breeding.

Improvement of hybrids

Thoday (1961) and his co-workers, using special tester strains in Drosophila have shown how genetic markers can be used to locate and evaluate linked loci affecting quantitative traits. Subsequently, Soller et al. (1976) concluded that relatively modest experiments, involving 1000-2000 F-2 or backcross individuals out of a cross between inbred lines, differing in one or more marker loci could enable differences in the quantitative value of chromosomal segments adjoining the differentiating markers to be determined. In fact, if the two strains differ in many markers, well spaced along the genome, a single F-2 generation will allow a total analysis of the entire genome, including main effects and dominance effects, for all quantitative traits scored.

The basic approach is shown in Table 1. It can be seen that a quantitative difference, \[ 2(1-2r)d \] will be found between the mean value of the two homozygous marker genotypes in the F-2. Similarly, the quantitative value of the homozygous marker genotype, expressed as a deviation from the mean of the two homozygous marker genotypes will be equal to \[ (1-2r)^2h \], and this parameter, divided by half of the main effect will be an estimate of \[ (1-2r)^2h/d \], the relative...
Table 1. Evaluation of main effect and dominance relations of quantitative loci in the F-2 of a cross between inbred lines differentiated at marker and quantitative loci.

<table>
<thead>
<tr>
<th>Mean value of genotypes at the quantitative locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
</tr>
<tr>
<td>Genotype</td>
</tr>
</tbody>
</table>

Cross between inbred lines differing at marker locus (M,m) and quantitative locus (A,a)

<table>
<thead>
<tr>
<th>Parental genotypes</th>
<th>MA</th>
<th>x</th>
<th>ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>+d</td>
<td></td>
<td>-d</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F1 genotype</th>
<th>MA</th>
<th>ma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>h</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>F2 genotype</th>
<th>MM</th>
<th>Mm</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>(1-2r)d+</td>
<td>(1-r)(2r^2)h</td>
<td>-(1-2r)d+</td>
</tr>
<tr>
<td></td>
<td>2r(1-r)h</td>
<td>2r(1-2r)h</td>
<td></td>
</tr>
</tbody>
</table>

Estimates of main effects and dominance at the quantitative locus

| Main effects | MM - mm = 2(1-2r)d |
| Dominance    | Mm - 1/2(MM + mm) = (1-2r)^2h |
| Relative dominance | Dominance = (1-2r) h^2 Main effects |

... dominance. The sign of the dominance estimate will give a notion as to the direction of dominance, but the magnitude of the estimate will be less than the actual dominance by an amount 2r. In favorable cases, some notion as to the value of r, and the absolute values of d and h can be obtained by maximum likelihood methods (Weller et al., to be published).

If the quantitative allele of interest is included between two closely linked marker alleles (a marker bracket say, MM-SS), then it can readily be shown that the expected mean value of the three homozygous marker genotypes in the F-2, MMSS, MmSs, and mmsS, will closely approximate d, h and -d, respectively. That is, in this case very good estimates of main effects and dominance effects at the quantitative locus can be obtained directly from the data.

Initial results in our laboratory (Weller et al., to be published) and elsewhere ( Tanksley et al., 1981; Zhuschenko et al., 1978) show that a wide variety of quantitative effects can in fact be found associated with specific genetic markers in crosses of this sort. Maximum likelihood methods show that
in some cases, at least, the results are due to linkage rather than pleiotropy. The results in our laboratory suggest that in tomato at least, the majority of marker-linked quantitative effects are of a magnitude that could readily be detected in experiments of about 1000 F-2 individuals.

Thus, the methods for genetic analysis described above can be utilized to identify specific marker-linked segments or brackets showing dominance or overdominance in the F-2 of a cross between two inbred lines. This would enable an existing commercial cross to be analyzed in terms of the number of chromosomal segments involved in producing the observed hybrid vigor, the type of genetic effect involved (i.e., complementary dominance, or overdominance), and the magnitude of effect of such segments. In addition, other inbred lines could be tested against the standard lines, and screened for the presence of additional dominant or overdominant segments. These could then be incorporated into the alternate standard line by marker-based introgression techniques, as described in Soller and Plotkin (1977) and Tanksley and Rick (1980). In this way chromosomal segments contributing to heterotic effects could be collected from a wide variety of resource lines and introduced into the standard lines, allowing direct and specific construction of lines showing maximum degrees of hybrid vigor.

Marker assisted selection in segregating populations

In species, such as dairy cattle, where very large progeny groups can be obtained from individual males, genetic markers or heavily marked chromosomal segments can also assist in within population selection. The basic procedure is aimed at evaluating alternative heavily marked homologous segments (denoted M1, M2) of a chromosome pair in a given male. Selection could then be carried out among future offspring of such an evaluated male on the basis of the estimated breeding value of the particular marked segments that the offspring inherited from his sire. Although the contribution of any single marked segment to genetic progress in this manner would be small, the joint contribution of a large number of such segments could be appreciable.

The procedure would be as follows. We denote the male parent as M1M2 and the female parents as MpMp, where M1, M2 and Mp are homologous chromosome segments, 20-30 cM in length and marked by 3 or more polymorphic markers. It is assumed that in the male parent, M1 and M2 differ at at least one of the marker loci included in the segment. The Mp are all of the corresponding homologous segments in the female parents, and may be similar or different than the M1, M2 segments. The cross of that male to a series of females will be of the nature M1M2 x MpMp, and the contrast HjTIp - HjFIp in the offspring will give information as to the quantitative value of the M1 and M2 segments in the tested male. The biometric equations defining the increase in breeding value of young sires chosen on the basis of the particular marker segments they received from their sire (following Soller, 1978) are given in Table 2.

As a particular marked segment is followed from a sire to his sons and evaluated again in the offspring of these sons, additional information will accumulate on the breeding value of this segment as compared to other homologous marked segments in the population. As information accumulates, future evaluations of the breeding value of alternative chromosomal segments in a sire will be based on past information as well as on new information obtained in the daughters of that sire. This will lead to more exact marker-based evaluations of young sires.

Eventually as this process continues the frequency of favorably evaluated segments will increase in the population to the point where essentially all segments carrying the favorable marker combination will have originated from the original marked segment in the originally evaluated sire. In addition, accumulated information will give a fairly precise relative evaluation of the various homologous marked segments segregating in the population. At this
Table 2. Biometric equations defining the increase in breeding value of young sires chosen on the basis of marker analysis of their sires.

\[ D = \text{observed phenotypic difference between daughter groups receiving alternate marker alleles or brackets from the sire.} \]

\[ \frac{2\sigma^2}{n} = \text{error variance of } D. \]

\[ \sigma^2_{AD} = 2pqd^2 = \text{genetic variance of } D, \text{ for a sire heterozygous at both ends of a marker bracket. } p \text{ and } q \text{ are frequencies of quantitative alleles, } \sigma \text{ is defined as in Table 2.} \]

\[ = 2pq(1-2r)d^2 = \text{genetic variance of } D, \text{ for a sire heterozygous at a single marker, with recombination probability } r, \text{ between marker and quantitative locus.} \]

\[ \sigma^2_D = \sigma^2_{AD} + \frac{2\sigma^2}{n} = \text{phenotypic variance of } D. \]

\[ b_D = \frac{\sigma^2_{AD}}{\sigma^2_D} = \text{regression of breeding value for } D \text{ on phenotypic value for } D. \]

\[ \bar{y} = \text{mean deviation from the overall daughter mean (for that sire) of those daughters receiving a particular marker bracket.} \]

\[ M = b_D(\bar{y}) = \text{expected breeding value associated with a particular marker or marker bracket.} \]

\[ B = \text{breeding value of dam of young sire.} \]

\[ I = \frac{1}{k} \sum_{i=1}^{k} M_i + B = \text{index value of young sire including breeding value of his dam and of } k \text{ evaluated markers.} \]

\[ \sigma^2_I = \frac{1}{k} \sum_{i=1}^{k} b_D^2 \sigma^2_{Di} + \sigma^2_B = \text{variance of index values.} \]

\[ i = \text{intensity of selection (Falconer, 1970).} \]

\[ S = \frac{1}{k} (\sigma^2_I - \sigma^2_B) = \text{increase in breeding value of young sire as a result of marker assisted selection.} \]

Point, the sires can be selected on the basis of their total genotype with respect to this segment. If a number of such segments are segregating simultaneously in the population, young sires would be selected on the basis of an index composed of the sum of their "known segment" values, and their dam breeding value. At this point marker assisted selection would be the major route to genetic improvement in the population. Table 3 gives biometric equations defining the increase in breeding value of young sires chosen on the basis of "known segments", and Table 4 gives a numerical example illustrating the increase in breeding value of young sires that might be achieved in this manner.
Table 3. Biometric equations defining the increase in breeding value of young sires chosen on the basis of "known segment" selection.

\[ a = d + (q-p)h \]  
\[ K_i = \text{breeding value of } i^{th} \text{ "known segment".} \]
\[ = 2qa, (q-p)a, \text{ and } -2pa \text{ for the } M_1M_1, M_1M_2 \text{ and } M_2M_2 \text{ "known segments"}, \text{ respectively.} \]
\[ I = \sum_{i=1}^{k} K_i + B = \text{index value of young sire including breeding value of dam and of all "known segments".} \]
\[ \sigma^2 \left( \sum_{i=1}^{k} K_i \right) = 2 \sum_{i=1}^{k} p_i q_i d_i^2 = \text{total variance of "known segment" breeding values.} \]
\[ \sigma^2 I = \sigma^2 \left( \sum_{i=1}^{k} K_i \right) + \sigma^2_B = \text{variance of index values of young sires, including dams breeding value.} \]
\[ S = i(\sigma_I - \sigma_B) = \text{increase in breeding value of index selected young as a result of "known segment" selection.} \]

Table 4. Numerical example illustrating the application of marker-assisted selection and of "known segment" selection of young sires in dairy cattle.

**Assumptions**

\[ \sigma = \text{phenotypic standard deviation of milk production} = 1000 \text{ kg.} \]
\[ k = \text{number of polymorphic marked segments} = 30. \]
\[ = \text{number of "known segments"} = 20 \]
\[ p = q = 0.5. \]
\[ d = 100, h = 0 \]
\[ n = 500 \text{ daughters evaluated per marker segment (1200 daughters per sire in all).} \]
\[ i = 2.0 \]

Then for:

**marker assisted selection**

\[ \frac{\sigma^2}{n} = 4000 \text{ kg}^2 \]
\[ \sigma^2_{AD} = 5000 \text{ kg}^2 \]
\[ \sigma^2_D = 9000 \text{ kg}^2 \]
\[ b_D = 0.56 \text{ for a marker bracket,} \]
\[ 0.44 \text{ for a single marker (r=0.1)} \]
\[ \sigma^2_B = 10,000 \text{ kg}^2 \]
\[ \sigma_I = 176 \text{ kg for a marker bracket,} \]
\[ 143 \text{ kg for a single marker} \]
\[ S = 150 \text{ kg for a marker bracket} \]
\[ 86 \text{ kg for a single marker} \]
III. COSTS

Costs per polymorphism determination

Three independent sets of activities are involved in demonstrating RFLPs. These are: (1) DNA extraction, (2) Preparation of probes, and (3) DNA analysis, including restriction, electrophoresis and blotting of DNA, and hybridization and autoradiography of the blots. DNA extraction is a major cost, coming to $15.00 per sample at current prices. Although a given DNA sample can be used for up to 240 RFLP determinations (20 slots per sample x 4 hybridizations per blot x 3 probes per hybridization), most applications would require far fewer determinations. Preparation of probes is a rather minor cost, coming to about $0.05 per RFLP determination. DNA analysis, again, is a major cost, coming to $0.30 per RFLP determination. The overall costs of an RFLP application would vary from $0.41 to $1.10 per polymorphism, depending upon the number of polymorphisms determined per DNA sample.

Costs per application

Table 5 gives estimated costs for the various RFLP applications described in the text, as a function of the number of DNA extractions required (equal to the number of individuals tested) and the number of RFLP determinations carried

Table 5. Costs for various RFLP applications, as determined by the number of individuals tested and the number of polymorphisms analyzed per individual tested.

<table>
<thead>
<tr>
<th>Application</th>
<th>Number of individuals tested</th>
<th>Number of polymorphisms per individual</th>
<th>DNA extraction</th>
<th>DNA analysis</th>
<th>Total ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial screening of population for RFLPs</td>
<td>20</td>
<td>1000</td>
<td>1200$</td>
<td>700</td>
<td>1900</td>
</tr>
<tr>
<td>Initial mapping of RFLPs</td>
<td>100</td>
<td>200</td>
<td>1500</td>
<td>7000</td>
<td>8500</td>
</tr>
<tr>
<td>Genetic analysis of a cross between inbred lines</td>
<td>1000</td>
<td>50</td>
<td>15,000</td>
<td>17,500</td>
<td>32,000</td>
</tr>
<tr>
<td>Introgression of a quantitative allele</td>
<td>500</td>
<td>2</td>
<td>7,500</td>
<td>350</td>
<td>7850</td>
</tr>
<tr>
<td>Marker assisted selection per evaluated sire</td>
<td>1000</td>
<td>50</td>
<td>15,000</td>
<td>17,500</td>
<td>32,500</td>
</tr>
<tr>
<td>Identifying a known locus$^a$ per locus</td>
<td>10,000</td>
<td>3</td>
<td>150,000</td>
<td>10,500</td>
<td>160,500</td>
</tr>
<tr>
<td></td>
<td>per 20 loci$^b$</td>
<td>40,000</td>
<td>600,000</td>
<td>840,000</td>
<td>1,640,000</td>
</tr>
<tr>
<td>Selection on the basis of 20 &quot;known-loci&quot; per individual</td>
<td>1</td>
<td>60</td>
<td>15</td>
<td>21</td>
<td>36</td>
</tr>
</tbody>
</table>

$^a$Based on $15 per DNA extraction, see Table 1. $^b$Based on $0.35 per RFLP determination, including $0.05 probe cost and $0.30 cost of DNA analysis, see Table 1. $^c$Based on 4 DNA extractions per individual to supply sufficient DNA for 1000 RFLP determinations. $^d$Based on 5 generations of backcrossing, 100 individuals typed per generation. $^e$Based on 1000 typed daughters per evaluated sire. $^f$See text. $^g$Assuming each individual tested contributed to evaluation of 5 marker-brackets. Cost of the entire program would be expended over 5-10 years.
out per individual tested. In most cases, costs are commensurate with the economic value of the breeding objective. For sire selection, costs are high, but in the case of outstanding proven sires whose sons will enter service as young sires, evaluation of the proven sire and marker-assisted selection among his sons might be economically feasible. The investments required to bring marker brackets to the stage of "known loci" seem high although they would be expended over a period of 5-10 years, but this application might become more attractive as experience accumulated on the value of marker-assisted selection among sons of outstanding sires, and on the magnitude of marker-linked quantitative effects in dairy cattle populations. Twenty "known loci" might well account between them for the bulk of genetic variation in milk production in dairy cattle, and their full exploitation over a relatively short period of time might yield eventual production increases on the order of 2000 to 3000 kg. per lactation. A return of this magnitude would be more than adequate repay of the investment involved.

REFERENCES


