Theor Appl Genet (1989) 77:95–101

Restriction fragment length polymorphism maps and the concept of graphical genotypes

N. D. Young* **and S. D. Tanksley**
Department of Plant Breeding and Biometry, 252 Emerson Hall, Cornell University, Ithaca, NY 14853, USA

Received May 24, 1988; Accepted June 29, 1988
Communicated by F. Salamini

Summary. With the advent of high density restriction fragment length polymorphism (RFLP) maps, it has become possible to determine the genotype of an individual at many genetic loci simultaneously. Often, such RFLP data are expressed as long strings of numbers or letters indicating the genotype for each locus analyzed. In this form, RFLP data can be difficult to interpret or utilize without complex statistical analysis. By contrast, numerical genotype data can also be expressed in a more useful, graphical form, known as a "graphical genotype", which is described in detail in this paper. Ideally, a graphical genotype portrays the parental origin and allelic composition throughout the entire genome, yet is simple to comprehend and utilize. In order to demonstrate the usefulness of this concept, graphical genotypes for individuals from backcross and F2 populations in tomato are described. The concept can also be utilized in more complex mating schemes involving two or more parents. A model that predicts the accuracy of graphical genotypes is presented for hypothetical RFLP maps of varying marker spacing. This model indicates that graphical genotypes can be more than 99% correct in describing a genome of total size, 1000 cM, with RFLP markers located every 10 cM. In order to facilitate the application of graphical genotypes to genetics and breeding, we have developed computer software that generates and manipulates graphical genotypes. The concept of graphical genotypes should be useful in whole genome selection for polygenic traits in plant and animal breeding programs and in the diagnosis of heterogenously based genetic diseases in humans.

Key words: RFLPs – Haplotypes – Selection

Introduction

The most widely used procedure for conveying information about an individual's entire genome is a cytological karyotype. In addition to showing the relative sizes and shapes of chromosomes, karyotypes can frequently be used to indicate the presence of inversion, translocation, or aneuploidy in a genome (Sharma and Sharma 1980). While the resolution of this technique can be high, as in *Drosophila* (Bender et al. 1983), in most higher organisms its resolution is poor, and only gross chromosomal properties can be characterized.

In addition to limited resolution, karyotyping is constrained by another, more serious limitation; namely, it does not generally provide information on the parental origin or allelic constitution of different regions of the genome.

In sexually reproducing organisms, chromosomes are mosaics of DNA stretches derived from preceding generations. In any meiosis, zero, one, or more crossover events may occur between a given pair of homologs. When crossover have occurred, a complete description of an individual's genome would include information on changes in allelic constitution due to recombination, as well as information on the locations where crossover events occurred. Although there are some cases in which this kind of information can be conveyed with traditional karyotyping (Tease 1978), in general, information on the origin of a given chromosome or chromosomal segment is not discernable from a simple microscopic karyotype.

By contrast, workers in the field of transmission genetics routinely describe individuals by their genotype at one or more genetic loci of interest. The description is generally alphabetic or numerical in nature, and it provides precise information on the derivation and allelic constitution at the specific loci. Until recently, however,

* To whom correspondence should be addressed

only a few loci could be followed simultaneously in most transmission genetic experiments.

High-density restriction fragment length polymorphism (RFLP) maps are now being developed in a wide variety of sexually reproducing organisms, including humans (Donis-Keller et al. 1987) and several crop plants (Tanksley et al. 1988; Burr et al. 1988). These maps can be used to determine the genotype of an individual at many, sometimes hundreds of loci, each defined by an RFLP marker. Thus, RFLP analysis makes it possible to deduce the most probable genetic constitution for regions throughout the entire genome in a given individual. However, as the number of loci analyzed in terms of RFLPs becomes very large, describing the numerical genotype at each and every locus becomes cumbersome and uninformative.

By contrast, portraying RFLP data in a graphical form would have a number of advantages over numerical genotypes. Such a "graphical genotype" would be similar to cytological karyotypes in describing an entire genome in a single graphic image, but different in that graphical genotypes would be inferred from RFLP data, and thus would show the genomic constitution and parental derivation for all points in the genome. In developing a graphical genotype, the primary goal would be to transform RFLP data, obtained in a numerical form, into an easily interpretable and accurate graphic image. In a practical sense, graphical genotyping based on RFLP analysis opens up the possibility of conveniently analyzing polygenic traits in the prediction of complex genetic diseases in humans (Lander and Botstein 1986) and in performing "whole genome selection" to breed for polygenic characteristics in plants and animals (Osborn et al. 1987; Tanksley and Hewitt 1988).

In this paper, we develop the mechanics for conveying RFLP data in the form of a graphical genotype, apply this new concept to backcross and F2 populations of tomato, and discuss several points relating to the potential power and application of graphical genotypes.

Materials and methods

Requirements for deducing graphical genotypes

In order to construct a graphical genotype, certain conditions must be met. First, a well-populated RFLP map for the entire genome must be available. In tomato, we have produced such an RFLP map by examining the restriction fragment pattern of several hundred cDNA and random genomic clones in individuals from both segregating F2 and backcross populations (Bernatzky and Tanksley 1986; Tanksley et al. 1988). This tomato RFLP map now consists of more than 300 markers on all twelve chromosomes (Fig. 1). Since many laboratories are currently developing RFLP maps for a wide variety of higher organisms, the utility of graphical genotypes may soon be realized for an ever wider range of eukaryotes.

In addition to a high density RFLP map, it is also necessary that the *cis-trans* configuration for the RFLP markers be known in order to prepare a graphical genotype. In populations derived from inbred lines, such as breeding programs consisting of backcross or F2 progeny, the *cis-trans* configuration can be inferred simply by knowledge of the breeding scheme. In more complex situations, complete RFLP data must be obtained for three generations in order to prepare graphical genotypes for individuals in the third generation. In humans, for example, RFLP data must be determined for grandparents and parents in order to develop graphical genotypes for the children in the pedigree. Without this knowledge of *cis-trans* configuration, RFLP data from some regions of the genome may have more than one possible graphical genotypes that are equally likely to be correct. In fact, an F2 population derived from selfing a true F1 hybrid may contain short genomic stretches that are intrinsically ambiguous (see below).

Assumptions employed in developing graphical genotypes

The primary assumption required for the development of graphical genotypes is that the genotype of a region between two RFLP markers is inferred from the genotypes of the markers that delimit the interval. When inferring the graphical genotype of an interval from the genotypes of the RFLP endpoints, there are often alternative configurations that will satisfy the available RFLP data. The rule is always to use the most likely configuration in developing a graphical genotype. Thus, simple configurations requiring the fewest number of crossover events are utilized in developing a graphical genotype, while alternative configurations that require one or more multiple crossover events are not. In practice, this means that if two consecutive loci have the same genotype, the genotype of the segment between the markers is inferred to be that of the two flanking RFLPs. When two adjacent loci have different RFLP genotypes, it is inferred that a crossover event has taken place somewhere between the loci.

Since the genotype of a non-recombinant interval is inferred from the genotype of its RFLP marker endpoints, double crossovers (or other even numbers of crossovers) in a given interval will falsify this inference, and the likelihood of double crossovers increases by the square of the probability of a crossover between the adjacent RFLPs. Thus, for any interval, the probability that the inferred genotype will be correct is:

$$\text{Probability} = (1 - p^2)$$

where p is the probability of a crossover event between adjacent RFLP markers. For the total genome, the probability that there are no incorrect intervals is:

$$\text{Probability} = \prod_{n=1}^{\substack{\text{total}\\ \text{intervals}}} (1 - p_n^2)$$

(These equations consider only double-crossovers and assume interference between crossovers to be negligible). As an example, consider an organism with a total genome size of 1000 cM in which RFLP markers are evenly spaced over the entire genome. Figure 2 shows the probability that a graphical genotype developed from such RFLP data is exactly correct (i.e., contains exactly zero intervals whose graphical genotype have been incorrectly inferred). Figure 2 also shows the expected proportion of the genome which is described correctly by the graphical genotype. This value is calculated by first determining the probability of 0, 1, 2, ... intervals that are incorrectly described for a given spacing of RFLP markers. These probabilities, along with the spacing size, are then used to determine the expected length of the genome correctly inferred, which is then divided by the total genome size (1000 cM in this example) to yield the expected proportion of the genome that is accurately portrayed by the

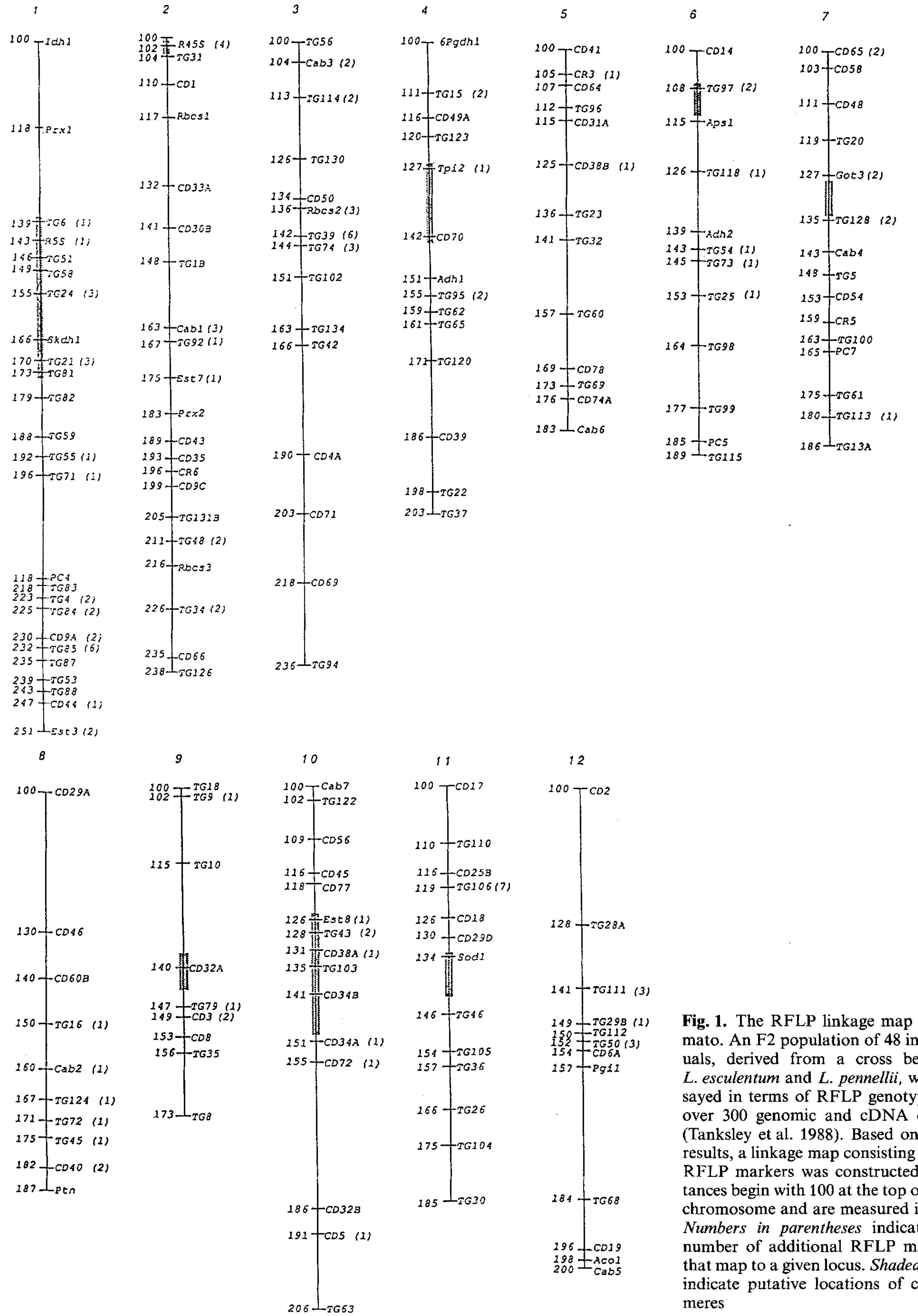

Fig. 1. The RFLP linkage map of tomato. An F2 population of 48 individuals, derived from a cross between *L. esculentum* and *L. pennellii*, was assayed in terms of RFLP genotype for over 300 genomic and cDNA clones (Tanksley et al. 1988). Based on these results, a linkage map consisting of the RFLP markers was constructed. Distances begin with 100 at the top of each chromosome and are measured in cM. *Numbers in parentheses* indicate the number of additional RFLP markers that map to a given locus. *Shaded areas* indicate putative locations of centromeres

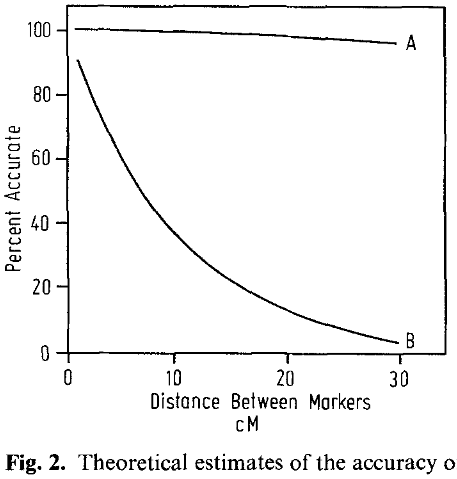

Fig. 2. Theoretical estimates of the accuracy of graphical genotypes. Curve *A* shows the proportion of a graphical genotype that is predicted to be accurate for a given density of RFLP markers in a hypothetical genome of 1,000 cM. Curve *B* shows the percentage of graphical genotypes with exactly zero incorrect intervals for the same hypothetical genome. Details are given in the text

graphical genotype. In Fig. 2, it can be seen that with RFLPs spaced every 10 cM, an inferred graphical genotype will have a probability of only 33% of being exactly correct for all regions (i.e., no incorrect intervals). However, this same graphical genotype will be accurate in describing the genomic constitution for over 99% of the genome. Even when the spacing between RFLP markers increases to 30 cM, the inferred graphical genotype will be accurate for approximately 95% of the genome.

Examples of graphical genotypes

Backcross populations, encountered frequently in plant breeding, represent one of the simplest situations for deducing graphical genotypes. In a backcross, progeny from a cross between two inbred lines are crossed back to one of the two parents (known as the recurrent parent). In tomato, we have constructed a backcross population involving two intercrossing species (*L. esculentum* and *L. chmielewskii*) and analyzed more than 200 plants for RFLP numerical genotypes at loci over the whole ge-

Plant 1:
1: 1111222222; **2**: 22221; **3**: 222222; **4**: 111122; **5**: 22222;
6: 21122; **7**: 111111; **8**: 11112; **9**: 2222; **10**: 2222222;
11: 112; **12**: 12222

Plant 2:
1: 1122222111; **2**: 11111; **3**: 111222; **4**: 111111; **5**: 22222;
6: 22211; **7**: 112222; **8**: 11111; **9**: 1111; **10**: 2222222;
11: 111; **12**: 21112

Plant 3:
1: 1111122222; **2**: 22111; **3**: 111111; **4**: 211111; **5**: 21111;
6: 22211; **7**: 222222; **8**: 22222; **9**: 1111; **10**: 2222111;
11: 111; **12**: 11112
a

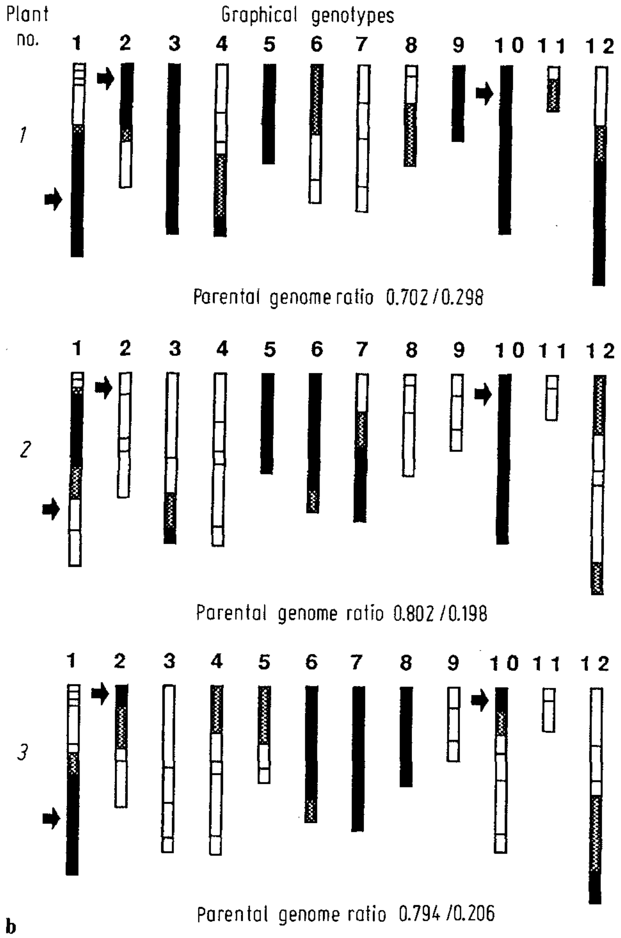

Fig. 3a and b. Graphical genotypes for individuals from a backcross population in tomato. F1 progeny from a cross between *L. esculentum* and *L. chmielewskii* were backcrossed to *L. esculentum* to produce a backcross population over 200 individuals (Paterson et al., 1988). Each of these plants was assayed in terms of 70 previously mapped RFLP markers. **a** Numerical RFLP data for three individuals in the backcross population, presented in order of the loci along the genome: 1 indicates homozygous for *L. esculentum;* 2 indicates heterozygous. **b** Graphical genotypes for these three individuals. *Intervals filled with white* indicate that the segment is derived exclusively from the *L. esculentum* parent. *Blackened intervals* indicate that the segment was derived from the *L. chmielewskii* parent, while *stippled intervals* indicate the presence of a crossover event. Note that a second homologue for each chromosome, consisting exclusively of DNA from *L. esculentum* (the backcross parent), is not shown in the graphical genotypes. *Arrows* point at regions controlling a hypothetical polygenic trait described in detail in the Discussion section. The parental genome ratio gives the proportion of the genome comprised of *L. esculentum* versus the proportion comprised of *L. chmielewskii*

1: 1122222222; **2**: 22222; **3**: 111122; **4**: 333333; **5**: 22333; **6**: 11111; **7**: 333332; **8**: 22111; **9**: 3333; **10**: 1233322; **11**: 111; **12**: 233333

a

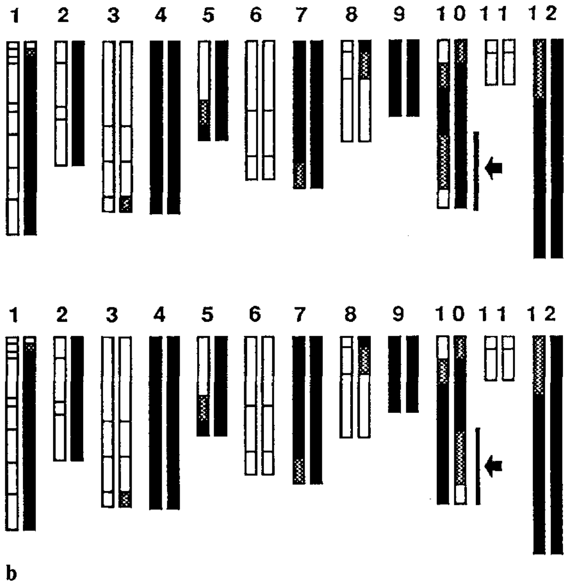

Fig. 4a and b. Graphical genotype for an individual from an F2 population in tomato. An F2 population was produced by selfing the F1 hybrid from a cross between *L. esculentum* and *L. pennellii* (Tanksley et al. 1988). The individuals in this population were assayed in therms of RFLPs for over 70 loci. **a** Numerical genotype data for one individual, presented in order along the genome: 1 indicates homozygous for *L. esculentum;* 2 indicates heterozygous; 3 indicates homozygous for *L. pennellii.* **b** Graphical genotypes derived from the numerical RFLP data shown in **a**. *White intervals* indicate segments derived from *L. esculentum, blackened intervals* indicate segments derived from *L. pennellii,* and *stippled intervals* indicate segments containing a crossover event. Both homologues of each chromosome pair are shown, since this is an individual from an F2 population. Two isomeric graphical genotypes of equal likelihood are shown that differ only in the region noted by the *arrow* and *thick line*

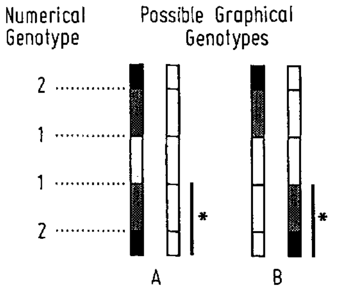

Fig. 5. *Cis-trans* ambiguity in F2 populations. A pair of isomeric graphical genotypes are shown to demonstrate the nature of the ambiguity found in graphical genotypes of individuals in an F2 population. Numerical RFLP genotypes are shown on the left and two equally likely graphical genotypes are shown on the right. These two isomeric graphical genotype differ in the *cis-trans* configuration of the segments that flank the homozygous region in the middle of each homologue. The region of uncertainty is noted by an *asterisk* and *thickened line*

nome (Paterson et al., 1988). RFLP results for three of these individuals are shown in Fig. 3a: 1 indicates that the plant is homozygous for the recurrent parent allele; 2 indicates that the plant is heterozygous. As noted earlier, when the number of loci examined is very large, this step yields a bewildering string of numbers, even if the data are presented in the linear order of the RFLP loci along the genome (Fig. 3a).

Numerical data, such as those shown in Fig. 3a can, instead, be profitably displayed in the form of a graphical genotype (Fig. 3b). Since these individuals are from a backcross population, only one of the two homologs is shown (the other homolog is derived exclusively from the recurrent parent, *L. esculentum*). The graphical genotype shows the complete genome, including the location of all of the RFLP markers examined, and conveys information on the parental origin of each segment of the genome for which RFLP data is available.

Also common in plant breeding are F2 populations derived from selfing an F1 hybrid. RFLP data for one individual from such an F2 population, derived from a cross between *L. esculentum* and *L. pennellii,* is shown in Fig. 4a, and the corresponding graphical genotype is shown in Fig. 4b. In the case of F2 populations, three genotypes are possible for each locus. In Fig. 4a, 1 indicates that the individual is homozygous for *L. esculentum,* 2 indicates that the individual is heterozygous, and 3 indicates that the individual is homozygous for *L. pennellii.* As noted above, such F2 individuals may contain genomic segments which are intrinsically uncertain. The individual portrayed in Fig. 4b contains an ambiguous region, so two isomeric graphical genotypes are shown. Ambiguities occur when heterozygous loci are separated by a stretch of one or more homozygous loci. In this situation, two equally likely graphical genotypes are possible that differ in the *cis-trans* configuration of the the flanking heterozygous regions (Fig. 5). However, calculations based on the Poisson distribution indicate that only 6% of a genome consisting of 10 chromosomes of 100 cM each will be ambiguous. In the *L. esculentum* × *L. pennellii* F2 population, in which the genome consisted of 12 chromosomes and a total of 1500 cM, the mean percentage of the genome that was ambiguous was approximately 9.2% (data not shown). Thus, the utility of graphical genotypes in F2 populations will not generally be seriously impaired by *cis-trans* ambiguities.

Discussion

Applications of graphical genotypes

The concept of graphical genotypes offers a more meaningful way to express the genetic composition of an individual than methods previously available. When more than a few loci are assayed with RFLP analysis, it becomes difficult both to express and comprehend an individual's genotype. Graphical genotypes reduce discrete locus data into a concise graphic image of an individual's linkage groups. These linkage groups are expressed in terms of the chromosomal regions inherited from the parents, reflecting chromosomal recombination and assortment. While genotypes have always been the product of such chromosomal recombination and assortment, in the past, they have been expressed largely independent of chromosomal karyotype. The advent of high density linkage maps based on RFLP markers now allows genotype and karyotype to be combined into an integrated graphical concept. In the near future, it is reasonable to expect that genetic linkage maps consisting of DNA markers will be correlated to physical maps derived from pulsed-field gel analysis (Smith et al. 1987). When this occurs, it will be possible to express graphical genotypes on the basis of actual physical segments of DNA, as well as on the basis of (recombination-based) map distances.

An obvious application of graphical genotypes is in the analysis of polygenic traits in animal and plant breeding programs. The main advantage provided by graphical genotypes is that they make it possible to perform "whole genome selection." In other words, graphical genotypes make it practical to comprehend, select, and manipulate the entire genotype of a given individual.

As an example, consider a hypothetical trait in tomato, such as resistance to some pest or disease, that is known to be controlled by several major loci located in known regions of the genome. Several examples of polygenic traits have already been described in tomato (Rick 1974; Zamir et al. 1984). The utility of graphical genotypes in such a situation is illustrated by Fig. 3 b. In this hypothetical example, three major loci controlling a polygenic disease resistance trait in tomato are noted by arrows beside the regions in which the controlling loci reside. Individuals with the genotype of parent B (blackened regions) at all three loci are resistant; however, other portions of the genome of parent B confer properties which are undesirable in other unrelated ways. The goal of a breeding program might, therefore, be to obtain an individual with the loci controlling disease resistance from parent B in a genome which is otherwise derived exclusively from parent A (white regions). This can be accomplished by repeatedly backcrossing progeny that have retained the resistance trait (i.e., retained the three determining loci) to the recurrent parent until the undesirable DNA from parent B is removed.

Comparison of the graphical genotypes of the three individuals shown in Fig. 3 b quickly indicates which individual in this hypothetical breeding program is most suitable for backcrossing. Among the three individuals, only plants 1 and 3 have retained the genotype of parent B at the three required loci. Apart from retaining the three resistance – determining loci, the most important goal of the breeding program is to remove unwanted DNA from parent 2. Comparison of plants 1 and 3 show that plant number 3 has experienced crossover events near all three of the target loci, thus removing large amounts of flanking DNA from parent B. In contrast, plant number 1 has retained large segments of DNA from parent B near all of the target loci. Moreover, the parental genome ratio, which is calculated by determining the proportion of the total genome derived from parent A versus the proportion derived from parent B, clearly shows that plant number 3 has less total DNA from parent B (20.6% DNA from parent B in plant number 3 versus 29.8% DNA from parent B in plant number 1). Thus, fewer total generations will be required to obtain an individual with the desired genome if plant number 3 is used for the next round of backcrossing.

Graphical genotypes can also be prepared for humans. Here, the primary goal of graphical genotyping would be to determine the likelihood of developing an inheritable disease that results from a complex genetic basis (Lander and Botstein 1986). As in breeding programs, the most important benefit of graphical genotypes would be in the ability to comprehend and interpret large amounts of numerical genotype data in a single graphical image.

Computer software for generating and analyzing graphical genotypes based on RFLPs

We have recently developed an Apple-Macintosh based program that converts numerical genotypes into graphical genotypes, as well as determines the parental genome ratio, for each individual in a backcross or F2 population derived from inbred parents. In addition, this program makes it possible to can carry out "selection" for a specific genome type in a segregating population based on parameters entered by the user.

Acknowledgements. The authors wish to thank M. Sorrells, A. Paterson, and E. Lander for their helpful suggestions on the manuscript. This work was supported by grants United States Department of Agriculture (85-CRCR-1-1609), National Science Foundation (DCB-8517773) and from a grant from the Cornell Biotechnology Program.

References

Bender W, Akam M, Karch F, Beachy P, Peifer M, Spierer P, Lewis EB, Hogness DS (1983) Molecular genetics of the bithorax complex in *Drosophila melanogaster*. Science 221:23–29

Bernatzky R, Tanksley S (1986) Toward a saturated linkage map in tomato based on isozyme and random cDNA sequences. Genetics 112:887–898

Burr B, Burr FA, Thompson KH, Albertson MC, Stuber CW (1988) Gene mapping with recombinant inbreds in maize. Genetics 118:519–526

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, Botstein D, Akots G, Rediker KS, Gravius T, Brown VA, Rising MB, Parkers C, Powers JA, Watt DE, Kauffman ER, Bricker A, Phipps P, Muller-Kahle H, Fulton TR, Ng S, Schumm JW, Braman JC, Knowlton RG, Barker DF, Crooks SM, Lincoln SE, Daly MJ, Abrahamson J (1987) A genetic linkage map of the human genome. Cell 51:319–337

Lander ES, Botstein D (1986) Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 83:7353–7357

Osborn TC, Alexander DC, Fobes JF (1987) Identification of restriction fragment length polymorphisms linked to genes controlling soluble solids content in tomato fruit. Theor Appl Genet 73:350–356

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature (in press)

Rick C (1974) High soluble-solids content in large-fruited tomato lines derived from a wild green-fruited species. Hilgardia 42:493–510

Sharma AK, Sharma A (1980) Chromosome techniques: theory and practice, 3rd edn. Butterworth, London

Smith CL, Econome JG, Schutt A, Klco S, Cantor CR (1987) A physical map of the *Escherichia coli* K12 genome. Science 236:1448–1453

Tanksley SD, Hewitt J (1988) Use of molecular markers in breeding for soluble solids content in tomato – a re-examination. Theor Appl Genet 75:811–823

Tanksley SD, Miller JC, Paterson A, Bernatzky R (1988) Molecular mapping of plant chromosomes. In: Gustafson JP, Appels R (eds) Chromosome structure and function. Plenum Press, New York, pp 157–173

Tease C (1978) Cytological detection of crossing-over in BudR substituted meiotic chromosomes using the fluorescent plus Giemsa technique. Nature 272:823–824

Zamir D, Selilabe-Davis T, Rudich J, Juvick JA (1984) Frequency distribution and linkage relationships of 2-tridecanone in interspecific segregating generations of tomato. Euphytica 33:481–488