

MultiMap: An Expert System for Automated Genetic Linkage Mapping

Tara Cox Matisse¹, Mark Perlin², Aravinda Chakravarti¹

¹Department of Human Genetics, University of Pittsburgh

²School of Computer Science, Carnegie Mellon University

Pittsburgh, Pennsylvania, USA

tara@chimera.hgen.pitt.edu, perlin@cs.cmu.edu

Abstract

With the advent of the Human Genome Project, the ability to rapidly construct comprehensive and accurate linkage maps based on genetic marker data from family studies is an absolute necessity. In addition to their usefulness in localizing genes for both simple and complex disorders, linkage maps are invaluable tools for genetic counseling using linked marker genes. Several computer program packages are publicly available which aid in the construction of linkage maps by computing multipoint likelihoods for specified locus orders. However, these programs work in a step-by-step fashion, requiring intensive user-intervention and analysis at each step. Such a repetitive process is amenable to computerized automation. We have developed and tested an expert system computer program, MultiMap, for automated genetic linkage mapping. This program greatly reduces the amount of user-computer interaction, increasing the accuracy and speed with which a map can be constructed. In addition, because the total mapping time is greatly reduced through automation, it is now feasible to explore and compare various mapping heuristics and mapping criteria in order to develop the most appropriate approach, or set of approaches, for genetic linkage mapping. MultiMap need not be restricted to the construction of genetic maps, but could be adapted to aid in the automated construction of physical maps as well.

Introduction

One of the primary goals of human genetics is to identify the genes which cause disease. Genetic linkage maps are extremely valuable for locating disease genes. These maps rely on the identification of genetic markers, or small segments of DNA that are known to vary slightly among individuals (i.e., DNA polymorphisms). Because the techniques involved in identifying and analyzing genetic markers have improved greatly in the past five years, at least 3,000 polymorphic genetic markers have now been identified in the human genome (Williamson et al., 1991). However, the computational process of constructing maps of these markers remains time consuming, and requires very specific expertise.

The first and most important step in constructing a linkage map is to determine the correct order of a set of

genetic markers. In order to guarantee identification of the single best order of a set of n markers, one would have to compute and compare the likelihoods of all $n!/2$ possible marker orders. As the number of markers increases, so does the total number of likelihood computations, as well as the length of time required for each computation. For example, even if one could calculate the likelihood of a particular order of twelve markers in 1 second, it would still take 7.6 years to evaluate all $12!/2$ (2.4×10^8) possible orders (Weeks, 1991). Thus, investigators rely on heuristic algorithms to determine a good, but not necessarily the best, order of loci in a reasonable amount of time.

There are several such heuristic approaches used to construct maps. These mapping algorithms follow a logical progression from beginning to end, with specific rules for each decision point. However, construction of a complete map often involves hundreds of steps, and the process of analyzing the map state and applying rules at each decision point is very time consuming and tedious.

The key benefit of our automation approach is the rapid construction of more accurate genetic linkage maps. Other advantages include more rapid and efficient map construction, as well as a mechanism for comparison of different mapping algorithms. By determining and replicating the control knowledge brought to bear at every decision making point in the map construction process, we have developed the MultiMap expert system. Our MultiMap implementation fully automates the specific heuristic approach which we have developed for constructing genetic linkage maps. Further, our detailed assessment of MultiMap on human chromosome 21 and on computer simulated data clearly demonstrates the accuracy, speed and cost effectiveness of this expert system approach.

Construction of Genetic Linkage Maps

As mentioned above, genetic linkage maps are very important for the identification of genes which cause disease. Typically an investigator first identifies on which chromosome a specific gene is located, determines the location of the disease gene on a map of genetic markers for that chromosome, and then isolates the gene itself through the use of molecular cloning techniques. Genetic counseling for a given disease becomes more

accurate as the location of the gene is delineated, and once the gene is isolated and cloned, the pathology of the disease can be understood and theoretically, better treatments and even a cure can be envisioned. For example, the genes for Duchenne muscular dystrophy (Monaco et al., 1986), cystic fibrosis (Rommens et al., 1989), myotonic dystrophy (Harley et al., 1992), as well as for several other disorders, have been identified through this approach.

Humans have 22 pairs of chromosomes (autosomes), plus two X chromosomes in females or one X and one Y chromosome in males. One member of each chromosomal pair is contributed by each parent. The particular DNA sequence at a given chromosomal locus is termed an allele, thus for each locus an individual has one maternal allele and one paternal allele. For many genetic loci there exist multiple alleles which have slight variations in their DNA sequence (polymorphisms). The more polymorphic an allele is, the greater its informativeness, or usefulness.

At each chromosomal locus, an individual has a 50% chance of having either one of his mother's two alleles at the same locus, and a 50% chance of having either one of his father's two alleles. Two genetic markers are said to be linked if the parental allelic combinations are preserved more often than would be expected by random segregation.

The degree of linkage, or genetic distance, between two markers is a function of the frequency of recombination. The measurement of genetic distance is expressed in map units or centiMorgans (cM, 1 cM = 1% recombination frequency). A maximum likelihood approach is used to determine the linear order, and inter-marker distances, of a set of markers. Several computer programs perform such maximum likelihood computations, including LINKAGE (Lathrop et al. 1985), MAPMAKER (Lander et al. 1987), CRI-MAP (Green, unpublished data, Lander and Green, 1987) and MENDEL (Lange, Weeks, & Boehnke 1988).

The goal of the Human Genome Project is to map enough genetic markers such that the average distance is 1-2 cM, with no gap being greater than 5 cM. Linkage analysis could then easily localize any gene to within 1 cM of a mapped marker. This leads to analysis of DNA segments of workable size, on which molecular techniques can be applied to isolate the gene.

Framework linkage maps consist of useful, evenly spaced markers whose order is well supported, although the distance between each marker may be quite large (low resolution). Framework maps may be used as a frame against which other markers can be mapped, creating a more detailed comprehensive map, with much finer resolution.

MultiMap Expert System

We have developed an expert system, MultiMap, which automates the construction of genetic linkage

maps. It begins with genotypic data for a set of markers and completes the construction of both framework and comprehensive maps without intervention by the user. It currently follows one specific mapping algorithm, developed in our laboratory, which has been shown to produce accurate maps (Cox, 1992). The details of this heuristic algorithm are presented in Figure 1 (Methods). MultiMap uses the computer program CRI-MAP for calculation of two- and multi-point likelihoods, and for estimation of recombination values. The user has control of 23 program parameters to help tailor each map to the specific data set being analyzed.

Initially MultiMap orders the markers by two measures of informativeness, so that more informative markers will be preferentially added to the map over less informative markers. Construction of a framework map then proceeds according to the given algorithm, with many decisions being made before the framework map is complete. For example, suppose there exists a map of three markers: A - C - D, with a distance of 25 cM between marker A and C, and 30 cM between marker C and D. Now suppose MultiMap finds that the marker B is located between A and C with significant odds. Because one of the framework map goals is that markers be evenly spaced, the sizes of the two new intervals created by adding B to this map between A and C must be analyzed. If B lies too close to either A or C, it would not be appropriate to add B to the framework map.

Markers are added to the framework map as in the previous example until either 1) the framework spans the entire length of the chromosome and all intervals are less than a user specified maximum size; or 2) no additional markers can be added to the framework map under the user specified conditions. At this point, MultiMap will expand the framework map into a comprehensive map.

The goal of a comprehensive map is simply to add as many markers as possible to the framework map. Thus, the algorithm for expanding a framework map into a comprehensive map is quite different from that used for construction of a framework map. For example, suppose there exists a framework map A - B - C - D, and an additional set of markers (E,F,G) are to be placed on this map. MultiMap may find that markers E, F and G all map to the same interval, between A and B. In this case, MultiMap must determine the order of E, F and G in the interval A - B. Thus, the likelihood of the following 6 marker orders must be calculated:

A - E - F - G - B - C - D	A - E - G - F - B - C - D
A - F - E - G - B - C - D	A - F - G - E - B - C - D
A - G - E - F - B - C - D	A - G - F - E - B - C - D

Once the likelihood of each of these six orders has been computed, MultiMap must determine whether any one order is significantly better than the other orders. At least 9 decisions are necessary to add these three markers to the map.

Because the order of markers in a map is somewhat dependent on the specific markers already in the map, the order of a map should be verified at several points during its construction. As in the above example, suppose we added the markers E, F and G to the map A - B - C - D. Assume the best order of E, F and G in the interval A - B was found to be E - G - F; this produces the map A - E - G - F - B - C - D. Because of the additional recombination information provided by the markers E, F and G, it is possible that the order of the markers B - C is now in an unlikely configuration. In other words, with E, F and G in the map, it may be the case that the order A - E - G - F - C - B - D is more likely than the previous order A - E - G - F - B - C - D.

Addition of markers to the comprehensive map and sequential order verification continues until no more markers can be mapped to single interval at the user specified level of significance.

There exist two other computer programs whose function is similar to that of MultiMap: the Gene Mapping System (Lathrop et al., 1988) and the BUILD function in CRI-MAP (Barker et al., 1987). However, neither of these programs is as flexible as MultiMap and both of these programs implement some restrictions which may result in an incorrect order. Thus, MultiMap fills an important niche in the process of identifying disease genes.

Methods

MultiMap is programmed to construct both framework and comprehensive genetic linkage maps following the heuristic algorithms described in Figure 1. These algorithms are described using pseudo-LISP notation.

MultiMap is written in the Common Lisp programming language (Steele, 1984). Lisp offers many features which make it especially useful and appropriate for the MultiMap program. The prototype version of MultiMap was developed for the Macintosh computer using the Macintosh Allegro Common Lisp package (MACL) (Apple Computer, 1989). This version took advantage of built-in Macintosh resources such as windows, menus and dialog boxes. The prototype also had graphical interfaces useful for displaying the framework and comprehensive maps. The present version of MultiMap is being run on a Sun SparcStation and has been tested with both CMU Common Lisp (Carnegie Mellon University, 1992) and under Lucid Common Lisp (Lucid, 1991). The Sun SparcStation has proven to be a very efficient platform for map construction.

While there were several programs to choose from for computation of multipoint likelihoods, MultiMap uses the CRI-MAP program (Green, unpublished data, Lander and Green, 1987). This choice was based on the fact that CRI-MAP is computationally faster and can analyze larger data sets than many of the other available programs. CRI-MAP is written in the language C

(Kernighan and Ritchie, 1978) and is easily ported to different computers.

A. Building a framework map.

1. Select the two most jointly informative markers.
2. LOOP UNTIL
(OR (AND
(all map intervals are smaller than a specified minimum,
the map is at least as long as a specified minimum))
no additional markers can be added to the map following these rules)
BODY:
IF (AND
(the next most informative marker can be placed in a single map interval)
(this marker would not be located too close to
neighboring markers))
THEN
add this marker to the map in the specified interval
3. Perform additional locus order verification to confirm the best order.

B. Comprehensive map algorithm.

1. Select a framework map.
2. LOOP UNTIL
No additional markers can be added to the map following these rules:
FOR every map interval
Determine map location(s) and likelihood(s) of all remaining markers
WHEN (any markers place to this interval only)
CONDITIONAL
((no other markers place to the same interval)
(add this marker to the map in this interval))
((all of the markers can be ordered within the interval)
(add all of these markers to this map interval))
(t
(add the single most informative of these markers to the map))
3. Perform additional locus order verification to confirm the best order

Figure 1. Algorithms used by MultiMap for construction of linkage maps.

Results

MultiMap has been tested on many sets of experimental data and on more than 20 sets of computer simulated data. The maps of experimental data were compared with published maps. The only difference detected between the maps produced by MultiMap and the expected maps was an inversion in the order of two neighboring loci. However, a simple local inversion of marker order is not considered significant. Each of the maps of simulated data was also compared with the expected results. Again, no significant discrepancies were detected in the order of markers in the maps constructed by MultiMap. The results of these comparisons are given below.

Experimental Data

We have constructed a linkage map of human chromosome 21 comprised of 43 PCR markers (McInnis et al. in press). Without the aid of automation as implemented by MultiMap, such a project could easily have taken a single two weeks. Instead, by using MultiMap, we completed this map in a few days.

A diagram of the comprehensive map of chromosome 21 constructed by MultiMap is shown in Figure 2. There were a total of 43 markers in the analysis which identify 39 genetic loci, 27 of which could be placed in single intervals with odds $\geq 1000:1$. A tick has been placed on the map at the location of each marker which could be mapped to unique locations, with the marker name printed to the left of the tick. The relative distance between the ticks represents the estimated map distance between markers. The interval locations of the remaining markers are indicated by bars to the right of the map. The total genetic distance covered by these markers on chromosome 21 is 66 cM. Thus, the average distance between markers is 2.5 cM. This average interval size is within the guidelines proposed in the Human Genome Project.

While the exact physical location for many of these markers is unknown, the order of the markers as determined by MultiMap agrees very well with the order published in other papers (NIH/CEPH Collaborative Mapping Group, 1992).

We have completed preliminary comprehensive maps of highly informative markers for a total of 23 chromosomes (22 autosomes and the X chromosome). These maps will be published in a forthcoming manuscript (Cox, Perlin, & Chakravarti, forthcoming).

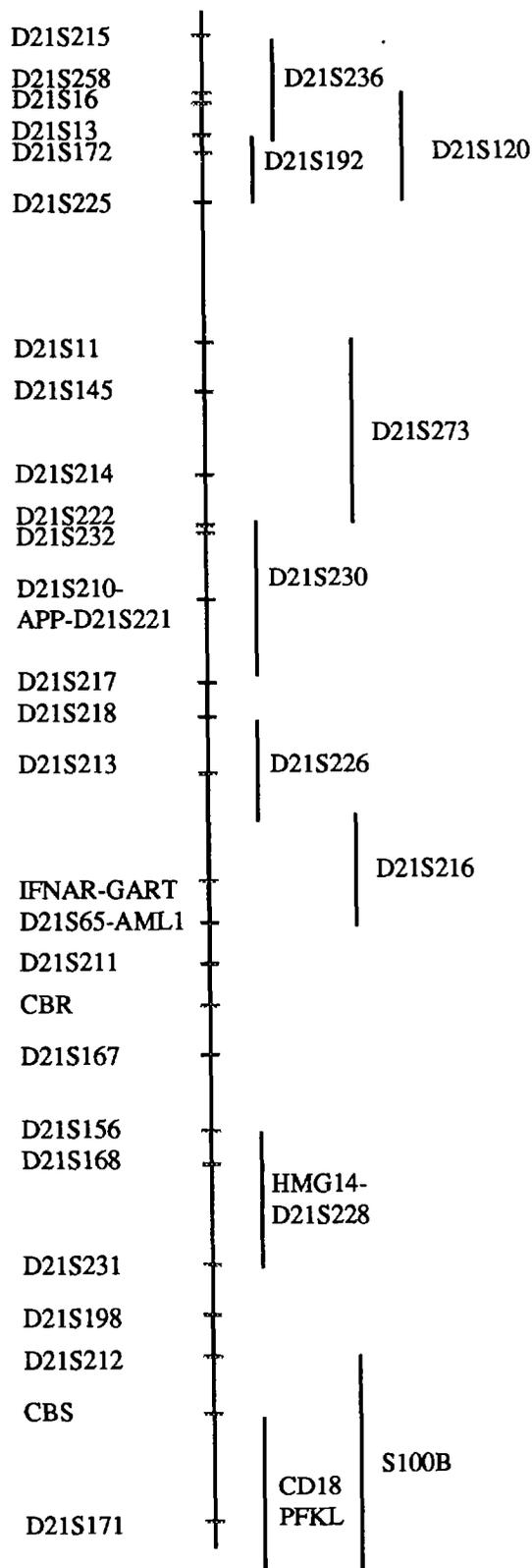


Figure 2. A genetic linkage map of human chromosome 21.

Computer Simulated Data

Genotypic data for over 600 markers in 20 different sets was simulated. The main difference between the simulated data and the experimental data is that the simulated data is error-free "perfect" data. This type of data is easy to generate and enabled us to greatly increase the number of tests performed on MultiMap. Each simulated set varied in the number of markers generated (11, 21, 51, 101), their average heterozygosity (30%, 70%, 100%), the distance between each marker (2, 5, 11 cM), and whether linkage phase was known or not. The ability of MultiMap to place markers in single intervals with odds $\geq 1000:1$ varied depending on these parameters. However, every marker that was localized to a single interval was found to map in the expected location. There were no instances of incorrect ordering of markers by MultiMap.

Discussion

The extremely high level of concordance in marker order between maps produced by MultiMap and comparable maps produced by other manual methods demonstrates the ability of MultiMap to produce accurate genetic linkage maps. Accuracy is one of the primary benefits of automation.

Accuracy is especially important in linkage maps, where errors are known to have drastic effects on the resultant maps (Buetow, 1991). Laboratory errors of several different types are commonplace in genetic linkage data, and have been shown to not only falsely inflate genetic map lengths, but to result in incorrect marker orders. While the effects of non-laboratory errors (i.e., data entry) on such maps have not been specifically studied, it is clear that both marker order and map lengths could be affected. When a map of 51 markers involving over 300 likelihood computations was constructed without the aid of automation, at least two typographical errors were made - giving an error rate of 0.67%. Such errors can be completely avoided through automation.

Obviously, another benefit of automation is increased efficiency. Computers can run 24 hours per day without fatigue. It is difficult to measure the amount of time saved through automation as implemented by MultiMap. For example, the NIH/CEPH Collaborative maps were produced by several different groups of investigators, each using his/her own personalized mapping algorithm, and each employing different numbers of computers and "mappers."

We can attempt to quantify the gain in speed incurred through MultiMap as follows: the construction of the framework and comprehensive maps of chromosome 3 took 13 hours and required 636 likelihood computations. The vast majority of the total time required by MultiMap is used for likelihood estimations, each estimated by CRI-MAP. When constructing maps without the aid of automation, for each likelihood

computation, a user must perform three actions: a) analyze the results of each likelihood computation; b) decide exactly which rule to apply to the current map state; and c) encode the appropriate parameters to perform the next likelihood computation. Assume that it takes a user who is fairly proficient with computers and has a good understanding of the mapping algorithm five minutes to perform these three actions. Over 636 computations, this user would use an additional 53 hours to complete the maps of chromosome 3, for a total of 66 hours. Conservatively, then, MultiMap can complete construction of a linkage map at least four times faster than a user who was applying the same mapping algorithm in a non-automated manner. Note that this assumes the human user must work continuously without breaks, and make no errors.

In addition to increased accuracy and efficiency, investigators can avoid the time consuming drudgery typically involved in map construction. They will no longer need to be tied to the computer for hours and days at a time. Because of the gain in speed through automation, different mapping algorithms can now be compared. Such comparison would have previously been cumbersome or impossible because of time constraints. There are several aspects of heuristic algorithms for map construction that have not been closely examined, but whose effect on the resultant map could now be clarified.

It would also be useful to allow the use of known physical location of markers when constructing linkage maps. For example, whenever a marker is to be added to the linkage map, its potential location could be compared with its physical location, when known, and this information could be factored into the resulting map.

In its present form, the MultiMap expert system is now an essential tool for construction of genetic linkage maps that is in routine use in our laboratory. We are currently incorporating additional knowledge sources to extend MultiMap for studying different approaches to mapping, and for comparing genetic and physical maps. In automating human decision making for genetic linkage mapping, the MultiMap project clearly demonstrates how expert system technology can eliminate a key cognitive bottleneck for a vital data analysis problem in molecular biology.

Acknowledgements

This work was supported in part by National Institutes of Health grant HG00344. Tara C. Maise is a postdoctoral fellow partially funded by the University of Pittsburgh, and is a fellow of the W.M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University, and the Pittsburgh Supercomputing Center. Figure 2 was drawn by Jillian Blaschak.

References

- Apple Computer, "Macintosh Allegro Common LISP," ver. 1.3.2, Macintosh Program, Apple Computer, Inc., Cupertino, CA, 1989.
- Barker, D., Green, P., Knowlton, R., Schumm, J., Lander, E., Oliphant, A., Willard, H., Akots, G., Brown, V., Gravius, T., Helms, C., Nelson, C., Parker, C., Rediker, K., Rising, M., Watt, D., Weiffenbach, B. and H. Donis-Keller. 1987. A genetic linkage map of 63 DNA markers on chromosome 7. *Proc. Natl. Acad. Sci. (USA)* 84: 8006-8010.
- Barnard, G.A. 1949. Statistical inference. *J.R. Statist. Soc.* B11:115-39.
- Buetow, K.H. 1991. Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am. J. Hum. Genet.* 49:985-994.
- Carnegie Mellon University Common Lisp (CMU CL), ver 16f. 1992. Carnegie Mellon University, Pittsburgh, PA.
- Cox, T.K. 1992. Automated genetic linkage mapping: development and performance assessment of an expert system computer program. Ph.D diss., Dept. of Human Genetics, University of Pittsburgh.
- Cox, T.K., Chakravarti, A. and M. Perlin. An expert system computer program for automatic construction of genetic linkage maps. 1991. *Am. J. Hum. Genet.* 49:366.
- Cox, T.K., Perlin, M. and A. Chakravarti. MultiMap: Automatic construction of linkage maps. 1992. *Am. J. Hum. Genet.* 51:A33.
- Harley, H.G., Brook, J.D., Rundle, S.A., Crow, S., Reardon, W., Buckler, A.J., Harper, P.S., Housman, D.E. and D.J. Shaw. 1992. Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* 355: 545-546.
- Kernighan, B.W. and D.M. Ritchie. 1978. The C programming language. Prentice-Hall Inc., NJ.
- Lander, E. and P. Green. 1987. Construction of multi-locus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. (USA)* 84: 2363-2367.
- Lander E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and L. Newburg. 1987. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1: 174-181.
- Lange, K., Weeks, D. and M. Boehnke. 1988. Programs for Pedigree Analysis: MENDEL, FISHER and dGENE. *Genet. Epidemiol.* 5:471-472.
- Lathrop G.M., Lalouel, J.M., Julier C. and J. Ott. 1985. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* 37: 482-488.
- Lathrop, G.M., Nakamura, Y., Cartwright, P., O'Connell, P., Leppert, M., Jones, C., Tateishi, H., Bragg, T., Lalouel, J.-M. and R. White. 1988. A primary genetic map of markers for human chromosome 10. *Genomics* 2: 157-164.
- Lucid Common Lisp/Sparc, ver. 4.0.0. 1991. Lucid Inc., Menlo Park, CA.
- McInnis, M.G., Chakravarti, A., Blaschak, J., and many others. A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics*, in press.
- Monaco, A.P., Neve, R.L., Colletti-Feener, C., Bertelson, C.J., Kurmit, D.M. and L.M. Kunkel. 1986. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* 323: 646-650.
- Morton, N.E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7: 277-318.
- NIH/CEPH Collaborative Mapping Group. 1992. A comprehensive genetic linkage map of the human genome. *Science* 258: 67-86.
- Rommens, J.M., Iannuzzi, M.C., Kerem, B-S, Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J.R., Tsui, L-C and F.S. Collins. 1989. Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping. *Science* 245: 1059-1065.
- Steele, G.L. 1984. Common Lisp: The Language. Digital Press, Burlington, MA.
- Weeks, D.E. 1991. Human linkage analysis: strategies for locus ordering. In: Advanced techniques in chromosome research. Edited by K.W. Adolph. New York: Marcel Dekker, Inc. p. 297-330.
- Williamson, R., Bowcock, A., Kidd, K., and many others. 1991. Report of the DNA committee and catalogues of cloned and mapped genes, markers formatted for PCR and DNA polymorphisms. Human Gene Mapping 11 (1991). *Cytogenet. Cell Genet.* 58: 1191.