Map-based cloning

Paper to read for this section :

Tanksley, S.D., Ganal, M.W. and Martin, G.B. (1995) Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. *Trends Genet.* **11**:63-68

The preceding chapter described methods used to use mutations to clone genes based on complementation and marker rescue. Both of those methods require some way to put DNA back into mutant cells. To screen or select for the desired clone in a recombinant DNA library, one not only needs a way to transform cells, one needs a way to transform at least as many cells as there are genes in the library (to determine the number of transformants you need to screen, use the Poisson equation). Unfortunately, both of these factors work against us as we move from microorganisms to higher eukaryotes. First, it gets harder to get large numbers of transformed organisms. Second, the amount of DNA that has to be examined gets larger as the genome complexity increases. Although tissue culture cells can be used to clone viral genes by marker rescue or complementation, these methods are basically impractical for isolating clones of genomic DNA from higher eukaryotes.

An alternative approach to using a mutation in order to get a gene is called map-based cloning. The basic idea behind map-based cloning is to clone the gene based on knowing its chromosomal location. In a map-based cloning, one starts with a genetic map of the organism's genome, finds a cloned marker that is close to the gene of interest, and then searches library DNA for clones that are near the previously cloned marker. By wandering around in the right neighborhood, one eventually clones the gene of interest. To understand the basis of map-based cloning, we will first review some common methods for generating genetic maps, and then describe methods for focusing cloning efforts on a particular location in the genome.

Making maps

A thorough discussion of genetic mapping is beyond the scope of this course. As in many genetic methods, a large number of very clever people have been working on ways to locate genes in the genome. The section below will only scratch the surface of what's in the literature.

©1999 J. Hu all rights reserved

Recombinational linkage mapping in eukaryotes

We’ve already discussed the idea that markers that are physically linked on the same DNA molecule can be separated by recombination during meiosis or mitosis. Mitotic recombination only results in a change in phenotype when a crossover occurs between two mutations in the same gene. Carlson observed mitotic recombination in strains heterozygous for different *suc2* alleles; the rates of recombination stimulated by UV light were used to construct a fine-structure map that ordered the mutations within *SUC2*. In metazoans, mitotic recombination is not a very useful method for mapping. First, the frequency of crossovers between markers in the same gene is very low. Second, somatic recombinants would have to be observed as the mosaics within an animal or plant.

Linkage mapping in higher organisms is largely based on the reassortment of markers during meiosis. The idea of genetic mapping dates back to the work of T.H. Morgan on *Drosophila* genetics around 1910. Morgan and his coworkers observed that the segregation some pairs of markers deviated from the statistical expectations for either independent assortment or complete linkage. This turned out to be due to the fact that the markers were on the same chromosome, but were occasionally reassorted by recombination. Where Carlson observed recombination by the formation of *SUC2* from two *suc2* alleles, classical genetics requires the ability to score single and double mutations among the offspring generated by gametes that are products of a meiosis where recombination may have occurred. As in the zebrafish mutant hunt, measuring recombination between recessive markers requires either inbreeding or tester strains where the mutations can be observed by their noncomplementation of markers in the tester strain.

Assuming that we can detect recombinant chromosomes, we can use the fact that the probability of a crossover between two genes is roughly proportional to their physical separation to construct a map of the affected genes. In his work on recombination in T4, Benzer noted that such a map provides two kinds of information, which he distinguished as topology and topography. The topology of the map simply gives us information about how the genes are arranged, i.e. we can determine that genes are arranged in a linear array without branching, and we can determine the order of the genes along that array. The topography is information about how far apart the genes are.

A more thorough discussion of mapping can be found in any introductory genetics textbook. Although we will not repeat those kinds of lessons here, I encourage you to review this material. Here, I only want to point out some of the take-home lessons about classical map generation.

Genetic mapping usually provides very good topological information; ambiguity tends to be a problem only for markers that are very close to one another so that recombination between them is rare. Topographic information is less precise for two reasons. First, the distances measured in units of probability of recombination don't always add up. Distances in genetic maps are expressed in units of centimorgans (cM), where 1 cM corresponds to 1% recombinant chromosomes among the gametes produced by meiosis from cells with a pair of markers. If the order of three genes is A B C, then A to B + B to C is often not the same as a direct measurement of A to C. Second, the assumption that recombination occurs randomly at all DNA sequences is clearly false. Mapping works because the local variations in recombination tend to average out over long stretches of DNA. Note also that recombination frequencies vary between species and even within species. For example, in *Drosophila*, there is no meiotic recombination in males.

The human map totals about 3300 cM. This does not mean that there are two markers that give 3300% recombinants; the maximum recombination that can be observed is 50%, which is what is observed for the independent assortment of markers on different chromosomes. Genetic distances of greater than 50 cM are generated by putting together all of the distances of a series of markers. Thus, if there are 20 genes that are evenly spaced at distances of 10 cM, the total separation is 200 cM. In this case, the most distant markers will segregate as if they are on different chromosomes because there will be, on average, two crossovers between them in every meiosis. However, their physical linkage can be inferred by the fact that each of them is linked to intervening markers that are, in turn, linked to one another. A set of markers that are linked in this way are called a **linkage group**. As the density of markers used to generate the map increases, the number of linkage groups will become the same as the number of chromosomes.

The resolution of mapping depends on two things. First, it is highly dependent on the number of offspring that can be examined. At distances of 1 cM, there will only be an average of 1 recombinant per 100 genomes examined. This is not a statistically significant number. High resolution mapping requires lots of offspring. Thus, while Benzer was able to map the rII region of T4 down to individual base-pairs, the resolution of mapping in higher eukaryotes is much lower. Second, mapping requires markers to map. High precision is not very useful at low marker density; if they are too far apart, markers on the same chromosome will behave as if they are unlinked.

Molecular markers

Traditionally, linkage maps have been constructed based on markers that give visible phenotypes in the organism, such as white eyes, altered wings or legs coming out of a fly’s head. The best maps were from well studied organisms like *Drosophila*, where there were lots of mutations, and where tester strains had been developed to make mapping easier. A lot of the elegance of classical *Drosophila* genetics lies in the clever use of combinations of point mutations, DNA rearrangements and fused chromosomes that simplify the detection of mutant phenotypes.

Mapping with markers that gave visible phenotypes was generally done using two or three mutations at a time, and worked best with nearly isogenic strains. This was because expression of a mutant phenotype could be affected by differences in the genetic backgrounds of parents and offspring.

Biochemical markers provided more allelic variation that can be used to generate genetic maps. Many mutations in the coding sequences of proteins are silent with respect to the function of a protein, but change the amino acid sequence in a way that can be detected. Enzymes with the same activity, but different sequences are called **isozymes**. If the differences are due to different alleles of the same locus, the variants are called **allozymes**. Different natural populations of the same species often express different allozymes. Often these can be distinguished by differences in electrophoretic mobility, especially on native gels where small charge differences will give dramatic differences in migration. Specific enzymes can often be visualized by staining the gels with substrates that change color when converted to products.

©1999 J. Hu all rights reserved

Allozymes increase the number of markers that can be used, and they are usually codominant. If one homozygous parent makes an enzyme that migrates at a different position from the enzyme made by a different homozygous parent, the heterozygous offspring will usually show both bands. Starting with heterozygous parents, a subset of the parental bands will be seen among different F1s (Figure 6-1).

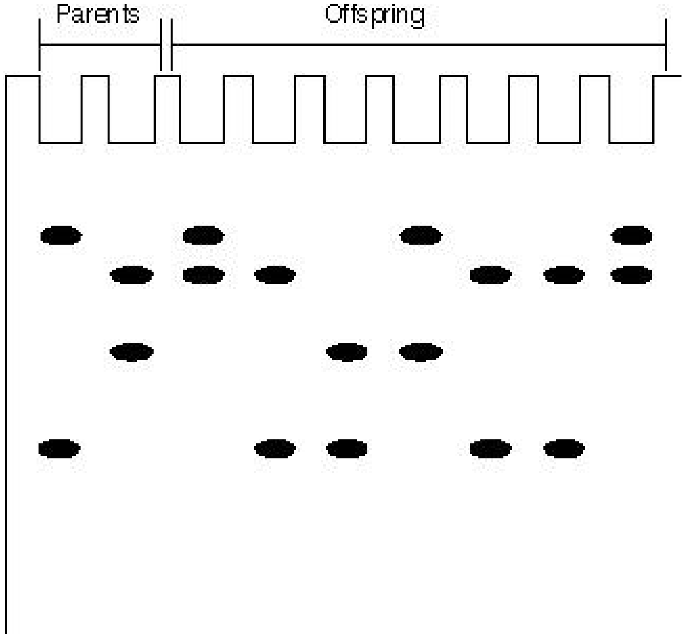

Figure 6-1. Segregation of allozyme markers among offspring of heterozygous parents.

Allozyme alleles will be codominant if the change in migration is due to a charge difference or a size difference in the primary sequence of the protein. Note however, that some differences in migration could be due to allelic variation in genes other than the one that encodes the enzyme. For example, a difference in the presence or absence of a certain glycosylation enzyme could alter the mobility of any protein that normally carries that posttranslational modification. Differences due to posttranslational modifications could lead to either dominant or recessive mobility differences, and some combinations of alleles might even create proteins with mobilities found in neither parent.

Although allozymes provide a useful source of allelic variation for mapping, the number of allozymes that can be examined is limited by our ability to find assays that will visualize the enzymes. For rapid mapping, allozymes have largely been replaced as a method of choice by DNA markers.

©1999 J. Hu all rights reserved

Variations in DNA sequences are ideal genetic markers. There is lots of allelic variation in natural populations; in fact, the variation is even more pronounced in intergenic regions where the need to conserve the function of a protein constrains the kinds of sequence changes that can occur. Like most allozymes, DNA sequences are codominant. Unlike traditional observable phenotypes or allozymes, the presence of different combinations of markers from different strain backgrounds does not affect whether or not the marker is observable, since you don't care at all about expression of the DNA sequence you are following. The important consequence of this is that you can follow as many markers in a set of crosses as you are able to resolve.

With so many advantages, why weren't DNA markers the first choice for generating genetic maps? Consider the historical context. Morgan and Sturtevant were using linkage to generate genetic maps in *Drosophila* more than 30 years before Avery and coworkers showed that DNA had anything to do with genes at all. Even after Watson and Crick's model showed in 1953 how the sequence of bases in DNA had to be important in the mechanism for heredity, detecting DNA polymorphisms was technically impossible until the late 1970s.

The first methods to detect subtle changes in DNA sequences were based on two technical breakthroughs: restriction mapping and Southern blotting. Restriction digests of genomic DNA from any free-living organism gives so many fragments that the DNA runs on a gel as a big smear. However, when the DNA is blotted to a membrane and hybridized to a radioactive probe corresponding to a specific sequence, a subset of the fragments are visualized as distinct bands. This method is called Southern blotting because it was invented by Ed Southern. Although the method has nothing to do with geographic directions, variations on the method have been named Northern, Western, Southwestern and Northwestern blotting.

The banding pattern for a particular probe varies among individuals. These variations are called **Restriction Fragment Length Polymorphisms** (**RFLPs**). Changes in fragment sizes can be caused by many different mechanisms: point mutations that create or destroy a recognition site for the enzyme, deletion of the site, deletion or insertion of DNA between two restriction sites, or insertion of DNA with additional sites.

RFLPs detected by Southern blotting provided the first DNA-based genetic mapping studies. One of the first outstanding examples of this method was done by Ray White and his colleagues at the University of Utah. White and his coworkers took advantage of the detailed genealogical records and large families from the Mormon population of Utah. This allowed them to follow the segregation of DNA markers through several generations, and to find molecular markers linked to genetic diseases.

Some high volume DNA marker methods

RFLP mapping was another leap forward in speeding the generation of genetic maps. However, detection of RFLPs by Southern blotting required probes for specific areas of the DNA. In addition, many DNA polymorphisms do not generate RFLPs. The development (or rediscovery, depending on which side of the Perkin-Elmer Cetus patent lawsuit you believe) of the **polymerase chain reaction** (PCR) by Kary Mullis in 1985 allowed the development of even more powerful methods for the identification of DNA markers. In this section I will describe two of the high volume PCR-based identification of DNA polymorphisms.

In the **Random Amplified Polymorphic DNA** (RAPD) methods, two or more arbitrary 10 mers with a 50% GC content are used as primers (Figure 6-2). Here, “arbitrary” means that each primer is a sequence picked at random with no attempt to match a genomic sequence. Each primer is a unique sequence; thus arbitrary primers are different from random primers, which are mixtures of different sequences. Under the appropriate annealing conditions, each primer will be able to prime DNA synthesis wherever there is a sequence that matches 3-4 bases at its 3’end. If two primers can prime synthesis toward each other within the distance that can be amplified by PCR (50-1000 bp using standard conditions), an amplified fragment will be produced.

Typically, primers and conditions can be used to amplify 50-100 fragments per reaction. Figure 6-2 shows how amplification can occur in a small region of the genome. In a real experiment, many more primer sites and fragments will be involved. The PCR products generated by the arbitrary primers are similar to the restriction fragments generated in an RFLP study in that their lengths can be affected by many of the same factors. Some of these are illustrated in the figure.

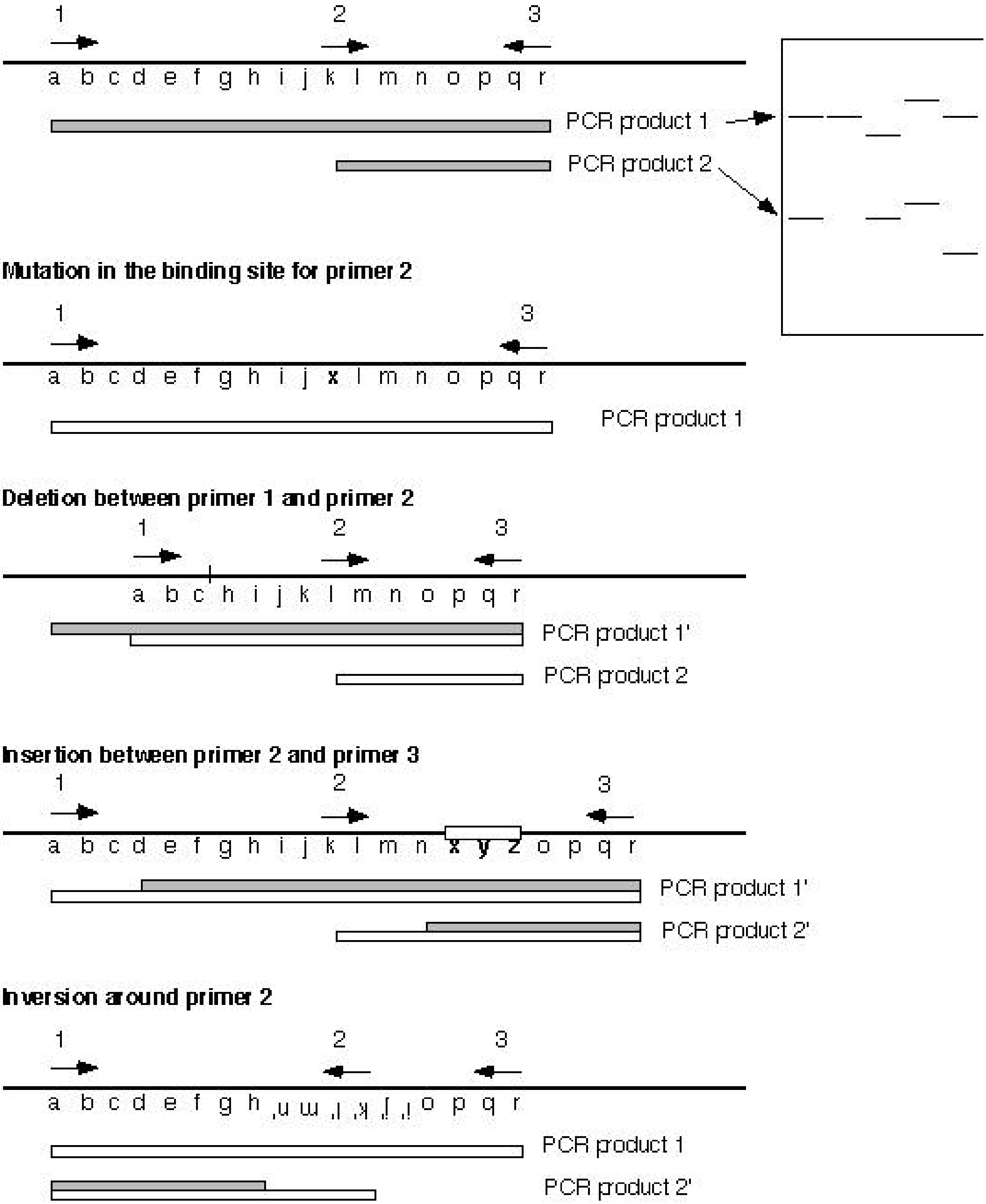
1
2
3
a b c d e f g h i j k l m n o p q r
PCR product 1
PCR product 2
Mutation in the binding site for primer 2
1
3
a b c d e f g h i j x l m n o p q r
PCR product 1
Deletion between primer 1 and primer 2
1
2
3
a b c h i j k l m n o p q r
PCR product 1'
PCR product 2
Insertion between primer 2 and primer 3
1
2
3
a b c d e f g h i j k l m n x y z o p q r
PCR product 1'
PCR product 2'
Inversion around primer 2
1
2
3
a b c d e f g h n m l k j i o p q r
PCR product 1
PCR product 2'

Figure 6-2 Detection of polymorphisms by RAPDs

©1999 J. Hu all rights reserved

To see a different set of amplified fragments, one just changes the sequences of the arbitrary primers used. Unlike RFLPs, the kinds of bands that can be produced by RAPD experiments is not limited by the sequences that are recognized by restriction enzymes.

Another PCR-based method for finding DNA polymorphisms is the **Amplified Fragment Length Polymorphism** (AFLP). Like RFLPs, AFLPs are based on restriction fragments. However, while RFLPs are found by probing a Southern blot with a specific sequence in order to visualize only those sequences that hybridize to the probe, AFLPs use a more generalizable way to visualize a subset of the total digest.

Figure 6-3 shows the critical components of an AFLP protocol. Genomic DNA is cleaved with two restriction enzymes. One enzyme is an infrequent cutter with a 6 bp recognition sequence, while the other is a frequent cutter with a 4 bp recognition sequence. Such a digest will generate thousands to millions of fragments and will run on a gel as a smear. Fortunately, we don't plan on running the digest directly on a gel.

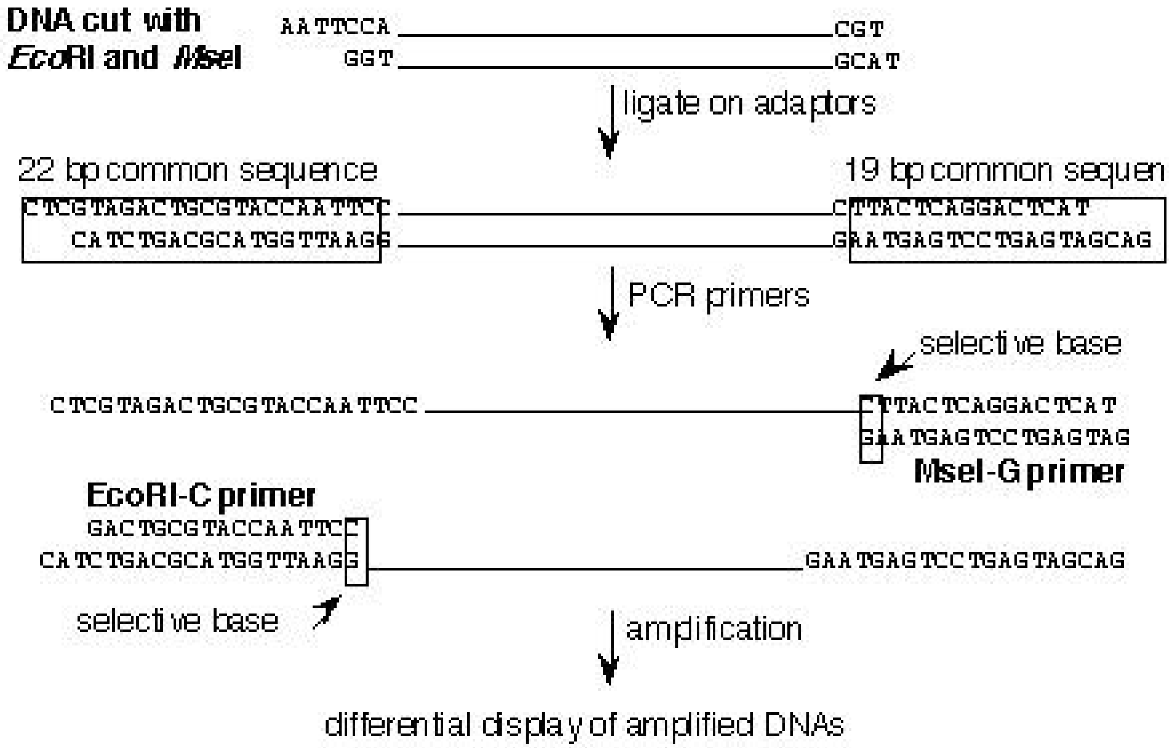

Figure 6-3. Outline of an AFLP protocol.

The two enzymes are chosen as having cleavage specificities that leave different overhanging ends. This allows us to ligate different adapters to each kind of end. We can then amplify the DNA with primers specific to each adapter. Fragments with the same kind of cleavage site at both ends (mostly 4 cutter fragments) will have the same adapter on both ends. During the melting and annealing steps in the PCR reaction, these molecules will form hairpins and will not be able to bind the PCR primers. They will be lost from the amplification.

This leaves us with PCR products corresponding only to fragments with the 6-cutter site at one end and the 4-cutter site at the other end. If we amplified all of these fragments, the number of different fragments would still be too large to resolve individual bands. However, we can selectively amplify only a subset of the fragments by using PCR primers whose 3' ends go beyond the adapter sequence. In the example shown in Figure 6-3, the primer corresponding to the adapter at the *EcoR*I end has an extra C at its 3' end. This primer will only be able to prime DNA synthesis on those fragments where the 3' C can base-pair with a G on the other strand. On average, this will eliminate 75% of the fragments. Similarly, the primer at the other end is also extended by 1 nt and will only prime 1/4 of the fragments. A PCR product will amplified only if both primers are able to prime DNA synthesis. Thus, only 1/4 x 1/4 = 1/16 of the sequences will be amplified. If this is still to complex to interpret, additional bases can be added to one or both primers to eliminate additional fragments. To see a different subset of the fragments, you just change the sequences at the 3' ends of the primers.

Both RAPD and AFLP methods provide good coverage of the whole genome without having to know anything about the sequence ahead of time. In both cases, the fragments generated are in a size range that can be resolved on sequencing gels. In addition, both methods can, in principle, be adapted for automated sequencing machines that detect fragments by fluorescent tags. Reactions done on two different strains can be marked with different fluorescent labels. Polymorphisms will show up as peaks that do not comigrate in the two fluorescence channels.

Markers generated by PCR methods can be used to generate genetic maps by the traditional method of calculating recombination frequencies between alleles. The density of markers that can be produced allows excellent maps to be generated. Note that having fragments is not enough, however. One needs a large number of markers that are different between two strains. Thus, DNA markers are most useful where matings can be performed between strains of the same species that have been separated long enough so that their genomes will vary at many locations. Chemical mutagenesis is not an efficient way to produce detectable DNA polymorphisms.

Linkage mapping with PCR-based DNA markers has been especially useful for genetic mapping in domesticated plants, where plant breeders have generated a wide variety of inbred strains that will be polymorphic over large stretches of their genomes. In addition, plants generate the large numbers of offspring that are needed to calculate map distances over short intervals.

Cotransduction and radiation hybrids

Linkage mapping requires parents with polymorphism, and it also requires either controlled breeding or large pedigrees. Both of these factors make linkage mapping especially difficult for humans. Another approach to mapping DNA markers has been developed that is conceptually similar to **generalized transduction** in bacteria.

Generalized transduction in *Salmonella* was described by Zinder and Lederberg in 1952. What they observed was the transfer of genes from one *Salmonella* strain to another. The transfer did not require physical contact between the cells; gene transfer would occur even if the cells were separated by a filter with holes too small for a bacterium to pass through. It turned out that the gene transfer was due to the fact that the donor cells were infected with a phage. Certain bacteriophages package their DNA in such a way that sometimes fragments of host DNA are incorporated into phage particles instead of viral DNA. These transducing particles can inject their host DNA fragments into a recipient cell, where recombination incorporates the DNA into the genome. Generalized transduction involves the packaging of DNA molecules of about the same length as the viral genome, where the DNA that is incorporated into transducing particles is a random sampling of the donor genome. Transduction is detected if recombination converts a marker in the recipient to the allele found at that locus in the donor.

©1999 J. Hu all rights reserved

Genes that are tightly linked can be packaged by the same phage particle. This is called **cotransduction**. Whether or not two markers can be cotransduced depends the distance between them and the packaging limit for the phage. For *Salmonella* phage P22, the packaging limit is about 40 kb, for the *E. coli* transducing phage P1, the limit is around 100 kb. Note that while the accuracy of distances inferred from cotransduction frequencies is not great, observing cotransduction at all means that the distance between the two markers is within the packaging limit.

Somatic cell genetics involves using tissue culture cells to map genes. For example, to map the human gene that encodes HGPRT, human cells can be fused to hamster cells in culture. With the appropriate setup, the interspecies hybrids can be selected. For example, one can start with a hamster cell line that has been selected to be *hgprt*$^-$, and introduce a gene for a drug resistance, such as hygromycin resistance. These cells can be fused to normal human cells, which will be *HGPRT*$^+$ and hygromycin sensitive. If the fused cells are grown in HAT medium containing hygromycin, then both parental types will be unable to grow (In practice, I’m not sure that a drug marker is needed to select against the human cells; one can use a human cell line that is unable to grow indefinitely in culture and fuse it to a hamster line that is already immortalized).

Hybrid cells tend to be **aneuploid**, which means they don’t have complete complement of chromosomes. Hamster-human cell lines tend to lose the human chromosomes. The human chromosome that contains the HGPRT gene will be retained, since it is required for viability. The chromosome can then be identified by looking at metaphase chromosomes in the microscope after treatments that give each human chromosome a distinct banding pattern. The analysis of chromosomes by this method is called **karyotyping**; it’s also used to examine fetuses for chromosomal abnormalities like Down’s syndrome. Note that individual hybrids will contain

This kind of somatic cell mapping was developed before PCR was available, and it works if there is a selectable marker on a human gene that you want to map. Note that we can use the same hybrid cell lines to do RAPD or AFLP analysis. When we compare the fragment pattern from the human parent to the human-hamster hybrid, anything that matches will map a DNA marker to the same chromosome as the selected marker. As in bacterial generalized transduction, only a subset of the genome is transfered to a recipient cell; markers that move together must be physically linked.

This idea can be extended even further and to much higher resolution by **radiation hybrid mapping** (Figure 6-4). Instead of using untreated human cells, one irradiates the human cells with X-rays prior to the fusion. This breaks the human chromosomes up into small fragments of a few hundred kilobases. After fusion to the hamster cells, some of the fragments will be incorporated into the hamster chromosomes. In the example shown, the human cells contain a functional thymidine kinase (*TK*) gene, the hamster cells are *TK*$^{-}$. The hybrid cells grow in HAT medium, while the hamster parent does not, because it can't use the exogenously supplied thymidine. The unfused parental human cells die due to radiation damage.

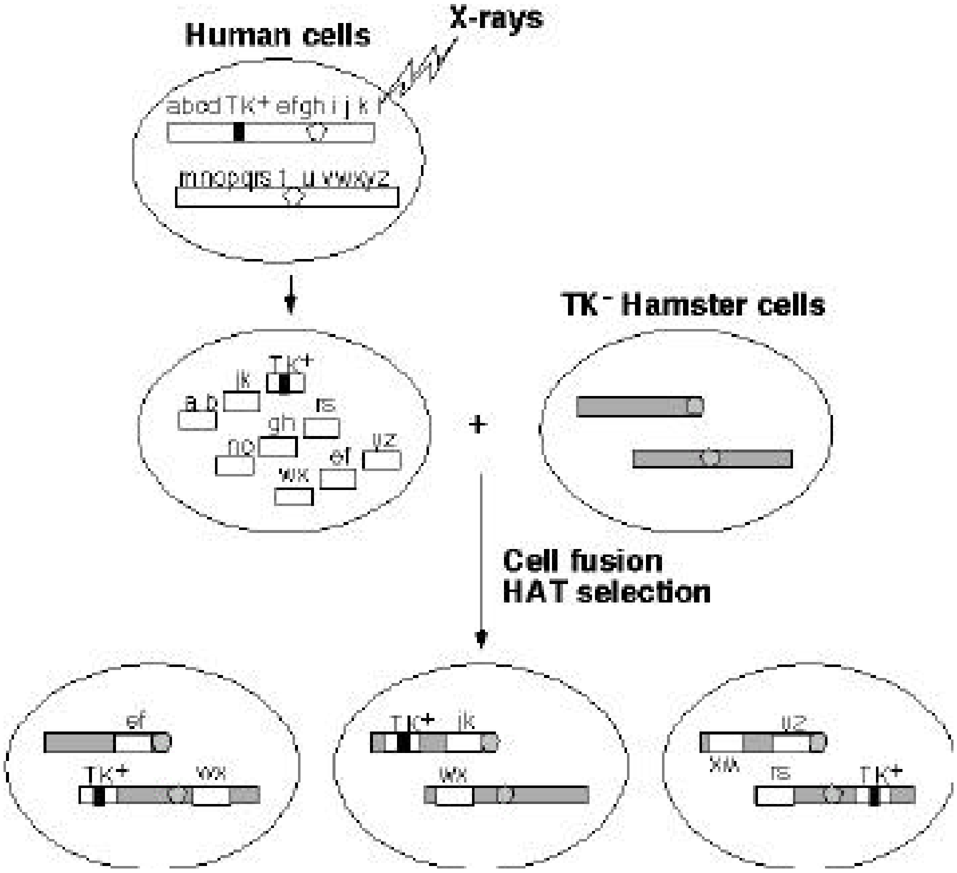

Figure 6-4. Radiation hybrid mapping

The individual clones are then analyzed by a high volume DNA marker method such as RAPDs or AFLPs or simply by hybridization with sets of cloned human cDNAs. Each individual clone will contain several markers, but the presence of two different markers in the same cell will be enriched if they are closely linked. By analyzing 100-200 hybrid clones, it is possible to map thousands of DNA markers. Note that radiation hybrid mapping allows placement of markers regardless of whether they are polymorphic in humans or not. This is especially useful for placing **expressed sequence tags** (ESTs; these are cDNA clones that are not necessarily assigned to specific genetic functions) on the map.

Radiation hybrid mapping was developed primarily for mapping the human genome, in order to overcome the difficulty of getting enough recombinant genomes for traditional linkage mapping. However, once the power of radiation hybrid mapping was demonstrated in the maps generated from human cells, it was realized that radiation hybrids could be used to supplement or replace mapping based on recombination in any organism. Radiation hybrids will be especially useful in the genome projects for domesticated animals, where breeding is possible, but is also slow and expensive.

Note that radiation hybrid mapping provides a good map of DNA marker, but that it can't completely replace linkage mapping. To figure out which DNA markers are linked to the mutation you isolated on the basis of a phenotype, you still have to use a mapping method that scores that phenotype. However, the DNA markers mapped by radiation hybrid methods can provide lots of landmarks in order to find a tightly linked marker, and as we shall see in the next section, we will need a linked DNA marker in order to perform map-based cloning.

Chromosomal walking

Now that we've examined some of the ways in which maps are generated, how do we actually do map-based cloning? The reading assignment by Tanksley et al. describes some of the available approaches. The problem can be divided into three parts: First, you need a way to clone genomic DNA that covers the region of the genome that contains your favorite gene. Second, you need a way to identify where the genes are in your cloned DNA. Third, you need to be able to determine which gene in the cloned interval is the one you want.

The standard method for map-based cloning, called **chromosomal walking**, was developed by Welcome Bender, Pierre Spierer and David Hogness at Stanford and was described in 1983 (J. Mol. Biol. **168**:17-33). Bender et al. developed walking and a related method called jumping to isolate DNA clones for three genes in *Drosophila*.

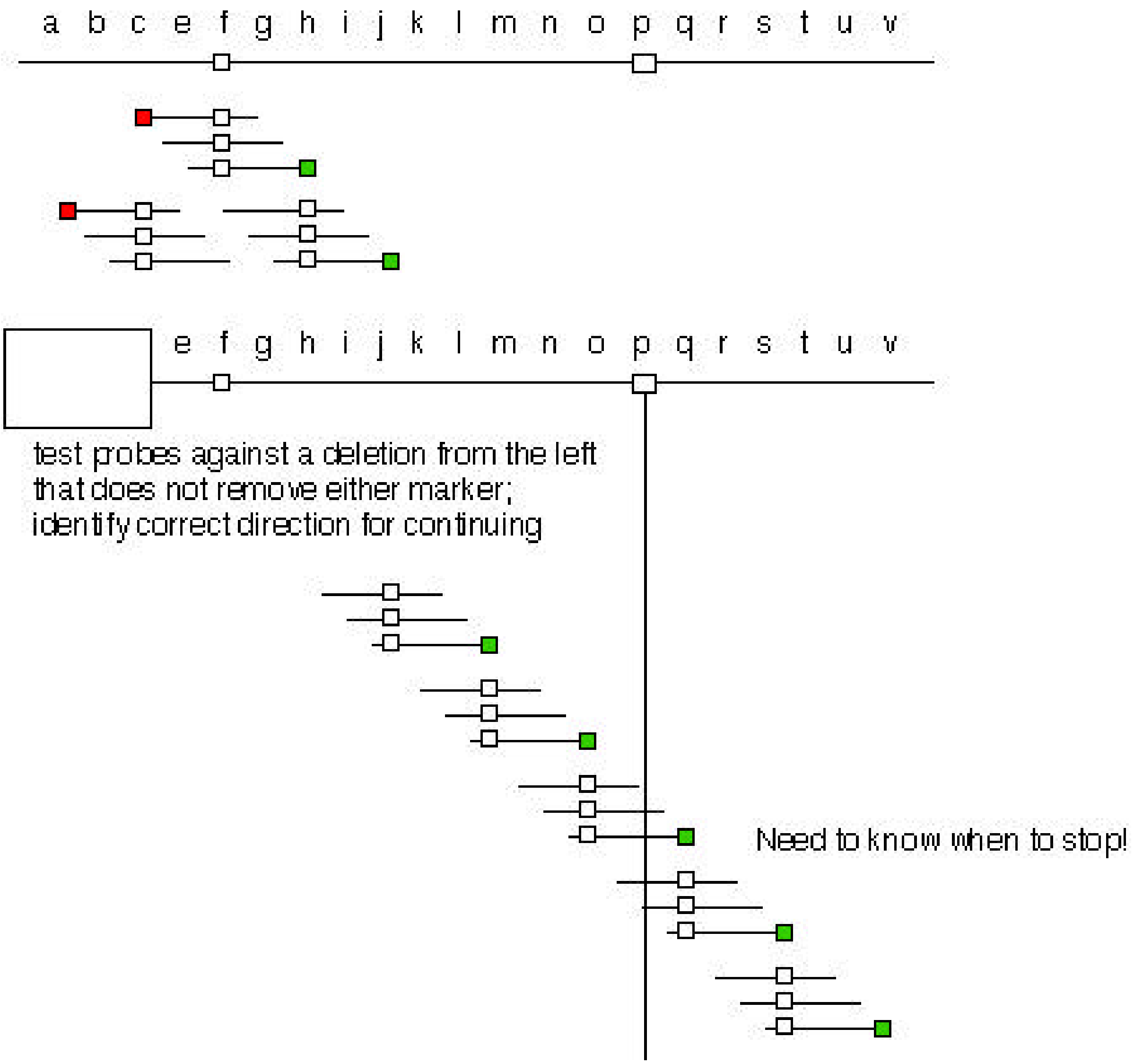
a b c e f g h i j k l m n o p q r s t u v
e f g h i j k l m n o p q r s t u v
test probes against a deletion from the left
that does not remove either marker;
identify correct direction for continuing
Need to know when to stop!

Figure 6-5. Chromosomal walking. See text for explanation

©1999 J. Hu all rights reserved

The general idea for chromosomal walking is shown in Figure 6-5. In order to start walking, you need a hybridization probe that maps near your gene. In the original *Drosophila* work, Bender et al found a cloned DNA that hybridized to the region of chromosome 3 that was known to contain the *rosy* and *ace* loci. Fortunately, they also found that the probe failed to hybridize to two deletion mutations that were known to be near or to cover *rosy* or *ace.* In general, the hybridization probe can be unique sequence shown to be linked to a mutation of interest.

Imagine that we have a probe that maps to position f and we want to clone a gene at position p. The first step in walking is to screen a genomic library and find all of the fragments that hybridize to the f probe. It is important to use a library that contains overlapping fragments; if you use a library from a limit digest with a single restriction enzyme, you won't be able to move in the walk. In the example, three clones were recovered (Bender et al. only found one that hybridized to their first probe). By restriction mapping and hybridization experiments, a map of the DNA covered by the three clones is assembled into a **contig**. From the map, we can determine which sequences are at the extreme ends of the contig and make probes to c and h. Using these probes, we go back to the library and isolate another series of clones. Assembling these into a larger contig with our original clones, we now cover the DNA from a-j.

Up to this point, we don't know whether the a end is going to the left and the j end is going to the right, as shown, or if the walk is going in the opposite orientation. In order to orient the walk, you want to use the terminal probes to test whether their hybridization can give you information about which way you are going. In the example I've made up, we have a deletion endpoint that we know comes in from the left. We also know that it does not take out our gene of interest at p. Hybridizing the a and j probes to DNA from the deleted strain, we can determine that the orientation of the walk is as shown.

Most large deletions remove essential genes and must be maintained in heterozygotes. Thus, we can't just extract DNA from the heterozygote and probe with the a and j probes; we'll see a signal from the wild-type DNA on the other homologous chromosome. Looking for a 2-fold difference in the

©1999 J. Hu all rights reserved

intensity of the hybridization between wild-type cells and heterozygotes is very difficult to do reliably. To determine whether a probe hybridizes to one or both homologs, it is preferable to use **in situ hybridization** methods, where the probe can be visualized on the chromosomes in the microscope. One way this can be done is to label the probe with a fluorescent tag. In our example, we could hybridize nuclei from cells heterozygous for the deletion with a mixture of the f probe labelled with fluorescein, which gives green fluorescence, and the a probe labelled with rhodamine, which gives red fluorescence (Figure 6-6). We'd expect to see nuclei with two green spots and one red spot, since only the wild-type chromosome will bind the a probe.

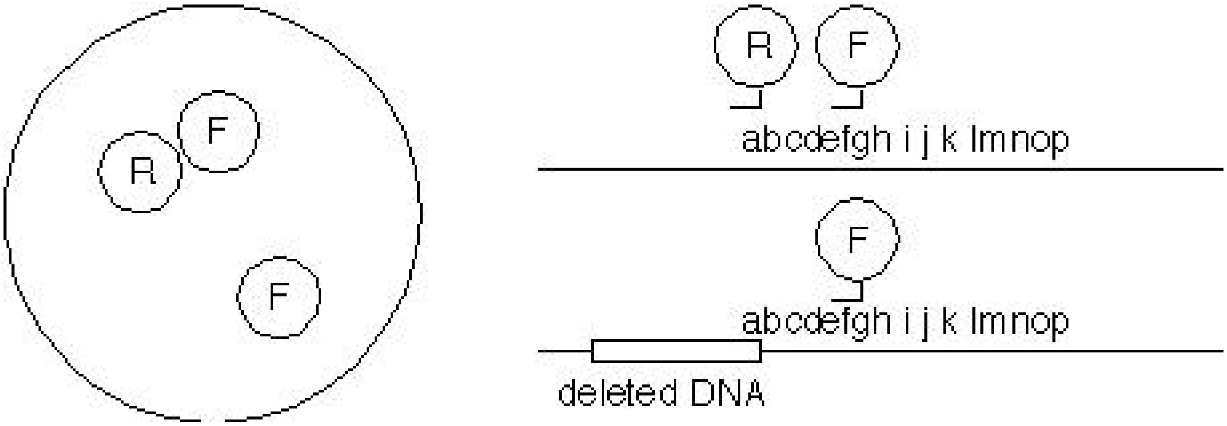

Figure 6-6 Fluorescent in situ hybridization with two probes. Left: cartoon of a stained nucleus. Right: How it's interpreted.

Knowing that the walk is going in the wrong direction with the a probe, we can use the j probe to continue the walk toward p. Each step in the walk consists of finding clones that hybridize to the rightmost probe from the contig, mapping the clones to extend the contig, and identifying DNA that can be used as a terminal probe in the next step in the walk. The cycles are repeated until you think you've passed the gene of interest. This can occur when you either make it to a deletion endpoint coming in from the other side, or when two walks started from opposite sides of the gene collide.

By walking, we can theoretically get from the starting marker all the way to the end of the chromosome. In practice, several factors limit walking. First, each step is relatively slow. The amount of DNA covered by each step in a walk is limited by the sizes of the inserts (Table 6-1)

©1999 J. Hu all rights reserved

Table 6-1. Insert sizes for different kinds of vectors

vector type	typical insert size
plasmids	1-10 kb
infectious λ vectors	10-20 kb
cosmids	35-45 kb
Phage P1 clones	70-100 kb
Bacterial artificial chromosomes	~ 300 kb
Yeast artificial chromosomes	100-2000 kb

Second, walks can end up teleporting to other parts of the genome if a terminal probe happens to contain repetitive DNA (Figure 6-7, top). This is relatively easy to detect with highly repetetive sequences, but can often lead to a dead-end roadblock. Even worse, if there are only two copies of the DNA corresponding to the probe, the walk could go awry without the error being noticed for a long time. Third, problems with the libraries can cause the walk to jump to the wrong part of the genome. A problem mentioned by Tanksley is chimerism (Figure 6-7, bottom). This occurs when a clone contains two different inserts ligated into the same vector. If a chimeric clone is isolated from the library by the ability of one insert to hybridize to the terminal probe, the second insert will always look like it's the far end of the contig. The new terminal probe will take you somewhere you don't want to go.

Chromosome landing

Tanksley's chromosome landing paradigm is based on two simple ideas. One can invest the effort in map based cloning either in the walking, or in finding the starting points. The first idea behind chromosome landing is to invest most of the effort into finding closely linked markers so that the first probe will "land" in the same clone as the gene of interest, or at most, one step away. Tanksley argues that high volume methods can almost always find one or more markers that are so closely linked as to make walking unnecessary, even though thousands of markers are needed.

The second idea is that one can find linked markers without bothering to make a map. Essentially, Tanksley is arguing that if the goal is to get a gene rather than to get a map, where the gene and the

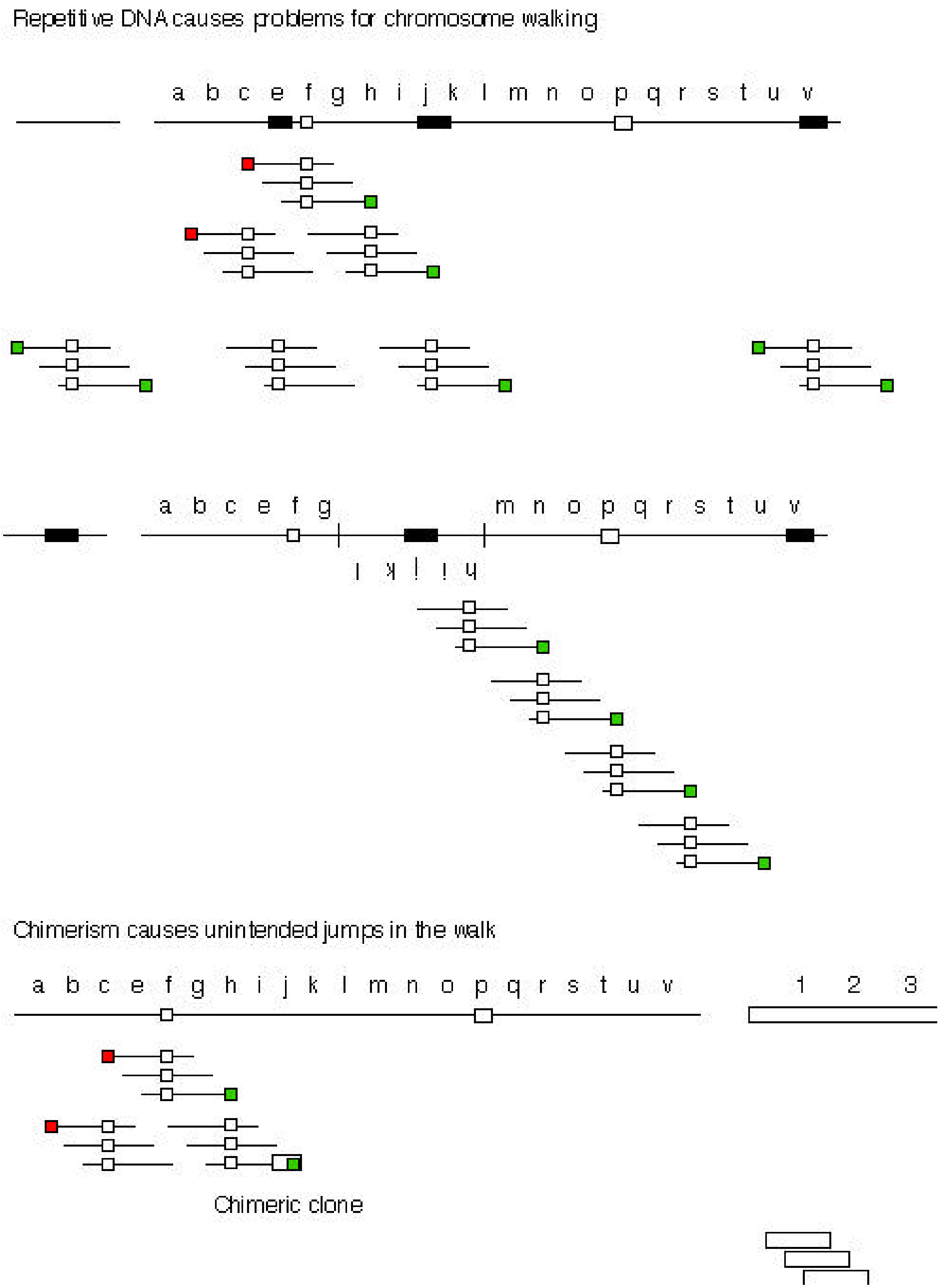
Repetitive DNA causes problems for chromosome walking
a b c e f g h i j k l m n o p q r s t u v
a b c e f g
m n o p q r s t u v
l k j i h
Chimerism causes unintended jumps in the walk
a b c e f g h i j k l m n o p q r s t u v
1 2 3
Chimeric clone

Figure 6-7. Problems that can occur during walking

linked marker lie with respect to the rest of the genome is irrelevant. Tanksley describes two approaches to finding a DNA marker tightly linked to a gene (See Box 1 in the Tanksley review).

The first method involves the comparison of nearly isogenic lines or strains (Figure 6-8). Nearly isogenic lines are strains originally derived from two parents that differ at many loci, including a gene of interest. One of the parents, P1, contains a *dominant* allele, *D*, that confers a desirable or interesting phenotype. Because the allele of interest is dominant, it can be detected in the F1. The F1 are crossed to one of the parents; the progeny of the first backcross are called the BC1 generation.

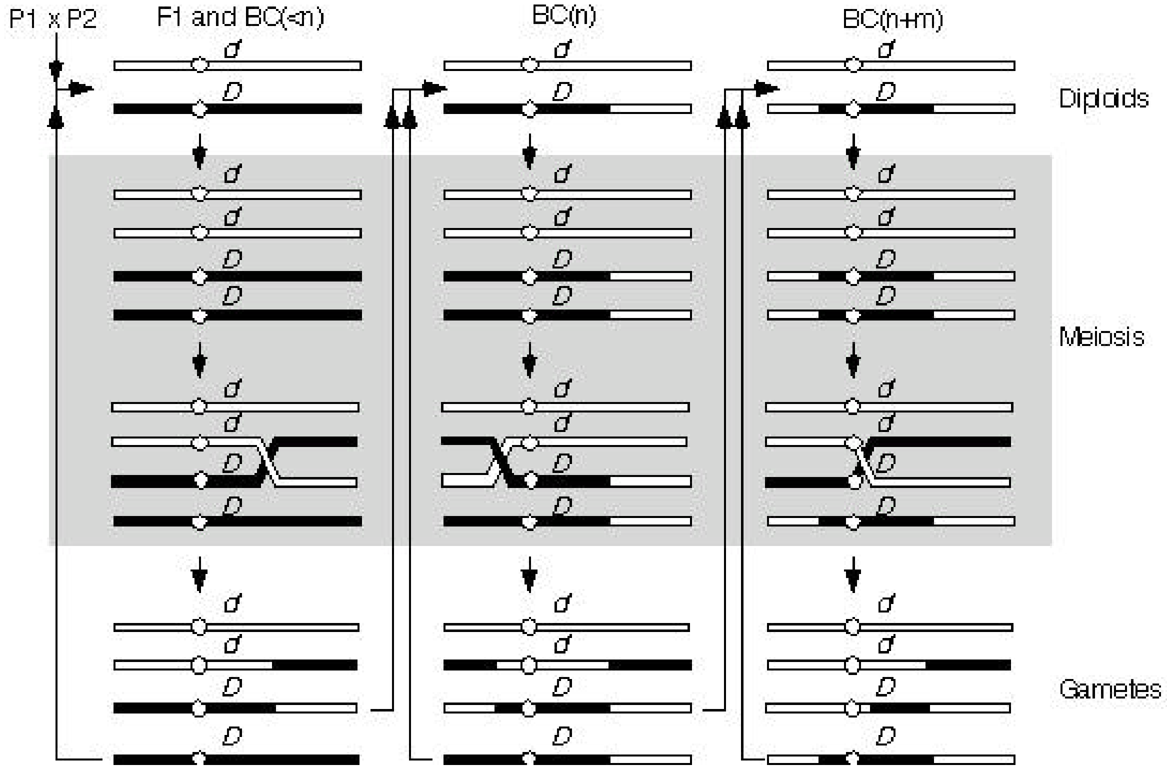

Figure 6-8. Generation of NILs. Only the chromosome carrying *D* is shown. Each column follows a generation through meiosis to the generation of gametes. If a chromosome that is nonrecombinant ends up in the next generation, the cycle within a column is repeated. If a recombinant chromosome is used for the next backcross, move to the right.

The F1 should be heterozygous at all loci. When the F1 goes through meiosis, each chromosome will assort independently, and the homologs will undergo recombination between homologs. Thus, the gametes from the F1 will contain different amounts of P1 and P2 sequences, and for an organisms with a large genome and significant recombination, the probability of inheriting a full complement of P1 genes is low. With each succeeding backcross, additional P1 DNA is lost from the line by assortment and recombination. However, *D* and markers closely linked to D are kept, because only the individuals with the *D* phenotype are kept for the next round of backcrosses. P1 markers will be progressively diluted out as the backcrosses continue. We can't predict exactly when each P1 marker will be lost from the line, but once it goes, it's gone forever. In addition, we can estimate the probabiliity that a given marker will still be there after a certain number of backcrosses as:

$(1\text{-}L)^n$

where L is the probability of a crossover occuring between the marker and the gene of interest, and n is the number generations of backcrosses. As Tanksley points out, generating a NIL from scratch takes a long time; after 10 generations, there is still a 60% chance that a marker 5 cM from the gene will still cosegregate with it. Using Tanksley's assumptions of a genome of 1000 cM and 10^9 bp, this narrows down the region of interest toabout 5 Mbp, an interval roughly the size of the *E. coli* genome! The NIL strategy is really only worth considering because generations of plant and animal breeders have already created NILs for many agriculturally useful markers by doing many generations of backcrosses.

The second method described by Tanksley is bulked segregants analysis (BSA). Starting with two inbred parents, a hybrid F1 generation is produced. As in the NIL case, the F1 is heterozygous at all loci that differ between the two parents. For BSA, however, instead of backcrossing the F1 to the parent with the recessive phenotype, one inbreeds the F1 and divides the F2 into two phenotypic classes. The class with the dominant phenotype will include individuals who are homozygous or heterozygous at all loci, including the gene of interest. As long as the dominant allele is provided from one parent, the individual will be placed into the dominant phenotypic class. With a large enough F2 population, the subpopulation with the dominant phenotype will contain individuals with every allele found in either parent.

The class with the recessive phenotype will be homozygous at the locus of interest. In addition, the subset of the F2 with the recessive phenotype will include individuals with markers from both parents. The probability of any given individual having a marker from the parent with the dominant allele will be proportional to

L x n, (XX what am I thinking here - check)

where L is the linkage and n is the number of individuals in the subpopulation. Thus, if we inbreed *Dd* F1s to give 4000 offspring, we would expect approximately 1000 DD homozygotes, 2000 *Dd* heterozygotes and 1000 dd homozygotes. Only the *dd* homozygotes would be included in the recessive subpopulation. Within these 1000 individuals, we would expect that one or both of the chromosomes carrying the *d* allele will have undergone recombination with its homolog from the *D* parent when the germ cells of the F1 went through meiosis. On average, we'd expect that a locus 1 cM from the gene of interest will be 99% from the d parent and 1% from the *D* parent, a locus 2 cM away will be 98% from the *d* parent and so on. For a *dd* population of 1000 individuals, There is a 95% probability of finding at least one individual with a marker from the *D* chromosome that lies within 0.3 cM of the gene.

In BSA, tissue from all of the individuals in each F2 subpopulation is pooled and subjected to a high volume marker analysis such as RAPD or AFLP. The pattern of bands that will be seen in the subpopulation with the dominant phenotype will be the sum of all of the bands seen for the two parental types. The recessive subpopulation will contain all of the bands from the *dd* parent and most of the bands from the *DD* parent. The small subset of *DD* bands that are missing in the recessive F2 population are tightly linked to the *D* allele.

This kind of analysis of pooled DNA would never work if we had to look for markers by Southern blotting or alloenzymes. In those cases, the intensity of the signal from a marker will be proportional to the abundance of the appropriate allelic form in the pooled material. Since the signal from all of the unlinked markers will be 50X stronger than the signal from a marker 1 cM from the gene, the latter will be indistinguishable from the background. In contrast, PCR can be easily set up so that at the end of series of cycles, the amount of a product from a marker found in 1% of the individuals in the population will be the same as the amount from a marker found in 50% of the individuals.

This paragraph and the next are my “hand-waving” explanation for why PCR can reduce or eliminate the differences in concentration of the original templates in a mixture. During a PCR cycle, priming occurs only if the primer can hybridize to an appropriate template strand. However, the primer is in competition for the template strand with the nontemplate strand that was melted off in the melting step. At low template concentrations, the primer usually wins. As the product accumulates, the concentration of the nontemplate strand will increase by up to a factor of two in each annealing cycle, while the primer concentration is slowly decreasing. This shifts the balance in favor of reannealling the two full-length strands.

In a mixture containing different starting amounts of DNA fragments, the abundant fragments will reach the point where rehybridization is favored first. Thus, as cycling continues, the amount of product from the template that was originally present in lower abundance catches up to the product from the abundant template.

Markers to clones by landing

BSA or NIL analysis will give us a set of DNA markers that are all tightly linked to the gene of interest. The bands corresponding to these markers can be cut out of the gel and reamplified to make pure hybridization probes, which can be used to find corresponding clones in a genomic library. An intermediate cloning step may be needed before generating probes, since many of the ways polymorphisms can occur are due to differences in the number of copies of a repeated sequence internal to the fragment. In order to find the right probe for sifting through the library, it may be necessary to use only a portion of the polymorphic marker that comprises unique sequences.

Note that even though BSA and NIL tell us nothing about the ordering of the markers, there is no need to map the markers relative to each other before cloning; it’s easier to assemble contigs of the clones than to order the markers by linkage analysis. Once the contigs have been assembled, the order of the markers will be apparent.

The probability of the gene being somewhere in the contig depends on how many tightly linked markers were found, and on the size of the inserts in the library relative to the linkage observed. If one only has one linked marker, the gene may be off the end of the clones isolated. With two markers,

©1999 J. Hu all rights reserved

there is a 50% chance that they flank the gene of interest. If this is the case, the gene has to be somewhere in the contig. As the number of linked markers increases, the odds of all of them falling on one side of the gene decrease.

The availability of methods that recover many linked DNA markers allows one to find all the clones that comprise a contig at the same time. It’s just a matter of doing several hybridization screens in parallel. Note that this is not possible with walking; one relies on finding the end of the contig at each step to make the probes for the next step in the walk. Thus, even if the contig generated by landing is just as big as the one that would have been generated by walking, you get there faster.

Finding the genes

Whether a contig is assembled across your favorite gene by walking, landing, or even by a genomic sequencing project, there is still the problem of figuring out where the gene of interest is within the contig. Tanksley describes typical contigs from plant genomes as containing about 30 genes. There are several approaches to figuring out which gene is the right one. As with the kind of cloning that can be used in the first place, what approaches can be used depends on what kinds of manipulations are technically possible for the organism under study.

Map based cloning is used when the efficiency of transformation is not sufficient to make cloning by complementation or marker rescue possible. Although it may not be practical to find the right gene in a library of 100,000 genes, checking 30 genes for complementation is not so bad, even if one has to go through the whole contig.

Another approach is to try to find the change in DNA sequence that corresponds to the mutation. Sequencing across the whole contig from both mutant and wild-type DNA is now possible, but it is usually more efficient to focus on areas that are likely to contain the gene. Genes can be identified by two kinds of approaches: hybridization and sequence analysis. Fragments within the contig that hybridize to cDNA clones must contain expressed genes. Note however, that it is difficult to be sure that all genes within a contig are represented in even the best cDNA libraries. When the complete sequence of the contig is known, there are computer programs that can be used to predict where the genes are with varying degrees of accuracy. Gene finding by sequence gazing is an ongoing area of **bioinformatics** research, but is beyond the scope of this class. For our purposes, I just want to point out that it is not enough to identify regions of sequence that have open reading frames. Even in prokaryotic genomes, there can be ambiguity about whether an entire open reading frame is used, and there are many genes that start with codons other than AUG. Eukaryotic genes are often spliced in complex ways, and intron-exon borders are not always obvious. Introns can be very large, and short exons can be difficult to distinguish from short ORFs that are not translated. Despite these problems, about 90% of the introns predicted in the *C. elegans* genome that were in regions that contained ESTs were predicted correctly, as judged by the existence of other experimental evidence.

Gene finding has focused on identifying genes that encode protein products. It is becoming clear that the variety of RNA gene products is much larger than had been previously suspected. In addition to tRNAs, rRNAs and small RNA molecules involved in splicing, there are RNA components in enzymes such as telomerase, RNase P, and the signal recognition particle that functions in protein secretion. RNAs that regulate the expression of specific genes have been identified in organisms ranging from *E. coli* to *C. elegans*, and RNA molecules have been shown to be involved in packaging DNA into some phage particles.

Regardless of how one identifies candidate genes, associating a sequence change with a mutation is not always straightforward. Since allelic variation in multicellular eukaryotes is often found by either heavy mutagenesis or from natural isolates, one has to worry about whether a sequence change in a gene is the mutation or if it's just a random polymorphism in the region. In an ideal situation, comparing different alleles in the same gene will give different sequence changes. Other information, such as the tissue-specific expression of the gene or its function as inferred from sequence comparisons, can be examined to see if they make sense in relation to the phenotype caused by the mutation. For example, it the genetics suggests that the gene of interest is involved in a signal transduction pathway, a gene that looks like a protein kinase would be considered to be a better candidate than a gene that looks like an enyzme in an amino acid biosynthetic pathway.

©1999 J. Hu all rights reserved