DEVELOPING MOLECULAR GENETIC MAPS

First molecular marker map: humans (Botstein et al., 1980)
Early plant mapping: tomato (Bernatzky and Tanksley, 1986), corn and tomato (Helentjaris et al., 1986), rice (McCouch et al., 1988)
Examples of saturated maps: tomato and potato (Tanksley et al., 1992), pepper (Livingstone et al., 1999), maize (Davis et al., 1999), rice (Harushima et al., 1998)

1. ***Identification of parents***

a. Must be polymorphic.

   i. Test for polymorphism by cutting DNA of potential parents with a number of enzymes assay polymorphism levels with several DNA markers.

   ii. Select parents with most polymorphism to cross or select the cross based on polymorphism.

b. Ideally of same species.

   i. Selfing species often have low polymorphism in the cultivated germplasm.

   ii. Interspecies crosses can have aberrant segregation due to genomic rearrangements.

c. Ideally inbred lines--often not possible with outcrossed species.

2. ***Identification of markers***

a. Must be easily scored–i.e., low copy number (repeated sequence probes give messy autoradiograms and the differentiating among loci is difficult)–and preferably codominant.

b. For RFLP–should screen out chloroplast probes.

3. ***Development of the segregating population***

a. A segregating population (F1, F2 or backcross) needs to be developed–this creates linkage disequilibrium that then enables recombination estimation. F2 or backcross populations are most commonly used, but F1 populations can be useful in highly heterozygous crops (e.g., alfalfa, sugarcane, potato).

b. <u>Population size</u> depends on how small a standard error you want for the recombination values (or how unambiguously you want markers to be placed on the map).

c. Many initial maps are made with 50-100 individuals–sufficient to make a map, but maybe not to use it, particularly for QTL analysis, where larger populations that provide higher resolution and more accurate location of genes/markers are needed.

d. Ideally, the mapping population will be immortal to permit continued DNA supply
<u>Means to permanence:</u>

   i. Vegetative propagation.

ii. Recombinant inbred lines (RI) formed by single seed descent from each F2–need to create the map in the RI population.

iii. Advanced generations of the F2 (i.e., F3 or F4) formed by bulking all progeny from selfing each F2, etc.

need to grow out a certain number of progeny and bulk leaf tissue to reconstitute the F2.

Note:

LINKAGE DISEQUILIBRIUM

Everything we've talked about in linkage is based on *linkage disequilibrium* (LD) within a population. LD is also called gametic phase disequilibrium. LD means that particular alleles at two loci occur together more (or less) often than expected by chance: e.g. A–B and a–b. If two loci are in linkage equilibrium, then the probability of allele "A" being paired with allele "B" equals the probability it is paired with allele "b."

The reason that we develop segregating populations, preferably from inbred parents, is that these crosses increase the amount of linkage disequilibrium, facilitating mapping. An F_1 population from the cross of two inbred lines is in complete linkage disequilibrium for loci with contrasting alleles between the parents. Thus, particular alleles are associated, so that when recombination is introduced (e.g., in the production of an F_2, backcross, recombinant inbred line, or advanced intercross line) we can detect linkage as the persistence of some linkage disequilibrium. The longer a population is randomly mated, the less LD will be present, and detection of linkage will become more difficult–i.e., shorter and shorter segments of chromosomes will remain in disequilibrium.

Mapping in outbred populations–e.g., a human population–can be done due to persistent linkage disequilibrium that had arisen at some piont in the past, most likely as a result of a population bottleneck. The population is probably not truly random mating and tight linkages will persist even after random mating, allowing the detection of linkage between a gene (e.g., cancer) and a molecular marker. However, for this to be successful, you need many DNA markers, which generally aren't available for most plant species. Several methods for LD mapping in human populations have been reported
(Jorde, 1995; Cheung et al., 1998).

Because of linkage disequilibrium, we can develop and use genetic maps in plants. The cautionary note, however, is that LD will dissipate under random mating, so markers useful initially in a population may lose their effectiveness if the population is intercrossed several times.

4. *Developing the map*

a. Screen probes on the parents; if polymorphic, screen on the segregating progeny population.

b. Score probes using some measure–e.g., A,H,B, or 1,2,3, etc.

i. for example:

				F2 progeny—>														
	P1	P2	F1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RFLP 1	A	B	H	A	B	A	H	H	A	H	B	H	H	H	B	B	A	A
RFLP 2	A	B	H	B	H	A	H	H	B	B	H	H	A	A	B	H	H	H
RFLP 3*	A	D	D	A	D	D	D	A	D	D	D	D	D	D	A	A	D	A

ii. *dominant marker–scored as D for presence (H or B) and A as abs.

iii. Note that parent 1 is always A and parent 2 always B.

c. Calculate χ^2 ratios for each marker to observe skewed segregation.

d. Analyze linkage between markers (as we have been discussing).

e. Placing classical genetic markers on map.
(Tanksley et al., 1992; Shoemaker and Specht, 1995)

i. any trait that is segregating in the population can be mapped. Score the trait in the same manner as a molecular marker (i.e., A for P1 and B for P2 or similar).

5. Saturating a map and targeting specific genomic regions

a. Specific regions may want to be targeted because they include a gene of interest or because they include few molecular markers.

b. Two Methods:

i. Nearly isogenic lines: for targeting regions with a gene of interest

a) Many plant breeding programs have developed NILs upon backcrossing a particular gene (e.g. disease resistance) into an otherwise desirable cultivar (e.g., Williams, Williams 79, and Williams 82 soybeans--the latter have phytophthora genes introgressed).

b) These NILs therefore have a putatively homogeneous genetic background, except in the location of the gene. One could screen molecular markers against these two lines, and any polymorphism that arose would be expected to be in the introgressed region.

c) Problems:

1) NIL are not readily available for all traits of interest (and are useless for saturating a molecular map).

2) Backcrossing is not an exact science: multiple genomic regions could still be remaining from the donor parent which would show polymorphism unrelated to the gene of interest.

3) The backcrossed segment may be very large so that markers detecting polymorphisms may be quite distant from the gene of interest.

ii. Bulked segregant analysis: (Giovannoni et al., 1991; Michelmore et al., 1991) for targeting regions containing a gene of interest (e.g., disease resistance) or which has few markers.

a) Use any population under study for which mapping data is available. Develop pools of individuals that are homozyogus for opposing alleles at a given locus or in a given region and screen the pools with molecular markers. Only markers closely linked to the markers used to make the pools will be detected (markers from other genome regions should be equally present in each pool). PCR-based markers (AFLP or RAPD) are most useful in this regard, because many potential markers can be assayed in a timely manner.

b) Develop two pools of individuals from within your segregating population:

1) For a single locus: pool 1 = individuals with the "A" allele for the marker
pool 2 = individuals with the "B" allele for the marker

c) Amplification products:

1) Differences can be distinguished clearly in a 10 cM window around a single marker and further away as differences in band intensity.

2) For amplification products that are far from the marker on which the pooling was done or on different chromosomes will be roughly even in intensity between pools--because the pool is heterozygous at all regions except that near the marker on which you pooled.

3) One problem is having slight amplification (or hybridization if using RFLP) if one or two individuals in the pool are the result of cross-overs between the target and the markers you are testing.

d) Numbers to pool: For a single locus, the more the better, but 10/pool is generally adequate

1) For a dominant marker unlinked to the target, probability of one pool having a band while the other does not is:

$$2\left[1-\left(\frac{1}{4}\right)^{n}\right]*\left[\frac{1}{4}\right]^{n}$$

2) So, for pools of 10 individuals, the probability of detecting a polymorphism between bulks unlinked to the target is 2 x 10^{-6}.

e) Bulks can be made for multiple regions at one time–i.e., for any genomic region of interest, two bulks can be constructed.

f) Bulks can also be made on a morphological trait–e.g., bulk on disease resistance, etc.

g) Can be used as a form of chromosome walking--identify a marker distal to one already mapped; then develop bulks based on the newly found marker and screen for linked markers again.

h) Confirming an amplification difference

1) Repeat PCR.
2) Hybridization of purified gel band to segregating population is necessary to prove that the marker is in the region desired (i.e., make the RAPD marker into an RFLP).

i) <u>Pooling based on two markers--interval targeting</u>

1) For an interval: pool 1 = individuals with “A” alleles at both loci flanking the interval.
   pool 2 = individuals with “B” allele at flanking loci
   i.e., Parent 1 contributes 'A' alleles at each locus, Parent 2 contributes 'B' alleles at each locus.

2) This method targets new loci to a particular interval rather than on either side of a marker

3) For an interval, many individuals per pool will increase the likelihood of a double recombinant, particularly if the interval is long–thus a tradeoff exists between false positives and no results due to double cross-overs. (see Giovannoni, 1991, p. 6556, Fig. 4).

j) Numbers of primers needed to screen:

1) Varies:
   i) In lettuce, 300 primers identified 3 markers linked to downy mildew resistance
   ii) In tomato, 100 primers–2 polymorphisms for chromosome 11; 200–1 on chromosome 10.

6. *When is a map done?*

The number of linkage groups should converge on the number of chromosomes as the map becomes “saturated”–that is, as all regions of the map are covered. Markers spaced every 5-10 cM is a good resolution for many applications. Saturated maps, like in tomato, have been developed with ~1000 markers and an average spacing <1.5 cM.

7. *How do I develop a high-density genome map?*

Selective mapping approach (Vision et al. 2000: Genetics 155:407-20) should allow us to develop a high-density global molecular linkage map. It has two steps.

1. Develop a reliable-framework map using a set of markers (framework markers), one marker representing each locus.
    a. Note that in this map you will find clusters of markers in most loci. The real challenge is to spread them into the intervals between loci of the framework map.
    b. Framework markers are used to classify a very large segregating population (in thousand(s) not in hundreds used in the framework map) into groups of recombinants. A group of recombinant progenies are those that carry a breakpoint (crossover or recombination event) between two adjacent framework markers. Thus, if there are n loci in i^{th} linkage group, then there should n-1 number of groups in that linkage group.
    c. Recombinants from two adjacent intervals (resulted from three adjacent framework markers) are then used to carry out the second phase as follows.
2. Two groups of recombinants from adjacent intervals can then be applied in mapping markers from a framework locus that distinguishes these two adjacent groups of recombinants.

Prospects and problems that may be encountered:

a. Only a subset of the original population is used in developing a high-density map -less work. From a segregating population of 1000 you will have only 10 individuals in a one cM interval.
b. Groups of recombinants can be utilized in hunting more molecular markers (as we have discussed in bulk segregant analysis).
c. Due to inhibition of recombination in heterochromatic regions we will fail to determine map positions of many markers that cosegregate. A regional mapping approach (screen more genotypes-thousands) may be applied if your gene of interest maps to that locus.

8. *Correlation of linkage groups with chromosomes*

a. rice: trisomics/monosomics (McCouch et al., 1988)
    i. Compare intensity differences among trisomics–hybrids vs. inbred trisomics.
    ii. Use two markers sequentially, or with hybrid trisomics, see allele differences easily.

b. *in situ* hybridization (Peterson et al., 1999)
    i. In this paper, they used single copy sequences as probes on synaptonemal complexes.
    ii. More commonly–use tandemly repeated probes to get stronger signal.

MAPPING F_1 and POLYSOMIC POLYPLOID POPULATIONS

I. Diploid or disomic mapping using inbred parents

Up to now, we've talked about controlled crosses between two inbred plants to produce an F_1 which segregates into an F_2 population. At any locus, we have a single allele from each parent that segregates in the population.

A similar method is used to map disomic polyploids–we just have two loci per marker if the plant is tetraploid--i.e. segregation of a marker independently in each genome. The only trick is to make sure you are scoring the appropriate alleles. In these crops, you develop a map for each genome–i.e., the number of linkage groups is the "n" number of chromosomes.

II. Mapping using non-inbred parents

For highly heterozygous crops, e.g., potato, alfalfa, and sugarcane, we can map in F_1 populations, because the parents (or at least one of them) has the possibility of carrying several alleles (two if diploid, four if tetraploid) that will segregate in the F_1 progeny. In these cases, recombination occurs within each parent.

A. F_1 diploid population (Ritter et al., 1990)

Assume that Parent 1 (P1) is heterozygous for two alleles, A1 and A2, and Parent 2 (P2) homozygous for one, A3. Further, we can say that A1 is on homologue 1 and A2 on homologue 2.

In a F1 progeny population, approximately ½ of the individuals will get A1 from P1, the other half A2 from P1, and all will have A3 from P2.

The most reasonable way to score these markers is to score A1 and A2 as two different loci for presence or absence of the band. No individual will have *both* A1 and A2, since they are diploids and can only get a single allele from parent 1.

		F1------>																				
P1	P2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A1A2	A3	13	23	13	23	23	13	23	23	13	13	23	13	13	13	23	23	23	13	23	13	23
A1		+	-	+	-	-	+	-	-	+	+	-	+	+	+	-	-	-	+	-	+	-
A2		-	+	-	+	+	-	+	+	-	-	+	-	-	-	+	+	+	-	+	-	+

Now consider a linked locus, B, with parent 1 having the B1 (on homologue 1) and B2 (on homologue 2) alleles and parent 2 having homozygous B3. If there is no recombination between A and B, the same F1 progeny that get A1 will get B1, and so forth for A2-B2.

B1B2	B3	13	23	13	23	23	13	23	23	13	13	23	13	13	13	23	23	23	13	23	13	23
B1		+	-	+	-	-	+	-	-	+	+	-	+	+	+	-	-	-	+	-	+	-
B2		-	+	-	+	+	-	+	+	-	-	+	-	-	-	+	+	+	-	+	-	+

In this case, we find that B1 will link to A1, but not to A2; and B2 links to A2, but not A1. We are defining linkage based on coupling only--what we are in effect building are homologue specific linkage groups.

We can link homologue groups for a specific chromosome together by looking for double alleles at two loci that each link together like this case-e.g., since both A and B have two alleles in P1, and since an allele of B links to an allele of A in both cases, we are fairly confident that we are looking at the same linkage group, but split into two homologues.

These alleles can be thought to be SDRF, or single dose restriction fragments. Indeed, in diploids, only SDRF or DDRF (double dose) are possible--a plant can either be heterozygous (in which case it has two SDRF) or homozygous (in which case it has a DDRF). However, a heterozygous plant may only have a single allele that can be mapped, because the second allele may be similar to one in the other parent, e.g. A1A2 and A2A2.

Thus, linkage groups can be built up for each homologue in each parent--in this case, two maps for P1, two for P2, or four overall.

Homologue groups can be linked between parents by looking at loci that segregate in each parent--e.g. C1D1/C2D2 in P1 and C1D1/C2D2 in P2, and develop a maximum likelihood expression for the observed numbers of progeny. (See Ritter et al. 1990. Genetics 125:645-654 for more info.)

B. **Mapping polyploids** (Grivet et al., 1996)

(Wu et al., 1992; da Silva and Sorrells, 1996)
Single dose restriction fragment (SDRF)–present on only one homologue
Double dose restriction fragment (DDRF)–present on two homologues

1. Consider SDRF first
An SDRF will segregate 1:1 in an F1 population (e.g. ½ of the progeny of A1A2A2A2 will have at least 1 A1 allele, not considering double reduction)

If an SDRF, coming from the same parent, is scored for each of two loci, then the expectations in the F1 are as follows (A=presence of marker 1; B=presence of marker 2):

Gametes	Coupling	Repulsion	Number obs.
A/B	$0.5(1-\theta)$	$0.25(1-w)+0.5w\theta$	a
A/-	$0.5(\theta)$	$0.25(1-w)+0.5w(1-\theta)$	b
-/B	$0.5(\theta)$	$0.25(1-w)+0.5w(1-\theta)$	c
-/-	$0.5(1-\theta)$	$0.25(1-w)+0.5w\theta$	d

where $w = 1/(h-1)$, h=number of homologous chromosomes (for a tetrasomic tetraploid, $w=1/3$)

Note that in this case, repulsion means that the loci are located on different homologues *of the same parent*.

Detect linkage using: $\chi^2 = (a-b-c+d)^2/n$, 1df

The maximum likelihood estimator of recombination fraction of two SDRF in coupling is:

θ (coupling) = $(b+c)/n$
θ (repulsion) = $[(h-1)(a+d)-0.5(h-2)n]/n$

where h=number of homologous chromosomes, n=total number of individuals in the population.
In repulsion, one homologue has a 1/(h-1) chance of ending in a gamete with another homologue. For the tetrasomic plant, each homologue has a 1/3 chance of going to the same gamete.

The bottom line is that only coupling markers are useful, because very large families are needed to see repulsion phase linkages.

E.g., alfalfa, 2n = 4x = 32

				F1----->											
Locus	P1	P2	1	2	3	4	5	6	7	8	9	10	11	12	Segregation
M1-1	+	-	+	+	+	+	+	+	-	-	-	-	-	-	1:1--SDRF
M1-2	+	-	+	+	-	-	-	-	+	+	+	+	-	-	1:1--SDRF

M1=Marker 1; two SDRF from parent 1 are scored (this means that P1's genotype is something like A1A2A3A3.

Since the two SDRF are both from P1, we could attempt to estimate linkage between them, but the solution ((4+4)/12) is outside the parameter space. This is because the SDRF, while from the same parent, are on different homologues and as such are not in coupling.

If we add a second marker, M2, which can be scored as a SDRF:

				F1----->											
Locus	P1	P2	1	2	3	4	5	6	7	8	9	10	11	12	Segregation
M1-1	+	-	+	+	+	+	+	+	-	-	-	-	-	-	1:1--SDRF
M1-2	+	-	+	+	-	-	-	-	+	+	+	+	-	-	1:1--SDRF
M2-1	+	-	+	+	+	+	+	-	-	-	-	-	-	-	1:1 (close)

If we did a linkage analysis here, we'd see that θ = 1/12 or 8.3%, indicating that M2-1 and M1-1 are on the same homologue.

In this manner, we build up linkage groups for each homologue, by only scoring markers from one parent, and testing recombination between the SDRF.

2. Linking homologue groups together

a. Using DDRF

M2=Marker 2; one DDRF is scored from parent 1–it's genotype is something like A1A1A3A3 and parent 2 is A2A2A3A3 and map A1A1 (M3-1).

			F1—>												
Locus	P1	P2	1	2	3	4	5	6	7	8	9	10	11	12	Segregation
M1-1	+	-	+	+	+	+	+	+	-	-	-	-	-	-	1:1--SDRF
M1-2	+	-	+	+	-	-	-	-	+	+	+	+	-	-	1:1--SDRF
M3-1	+	-	+	+	+	+	+	+	+	+	+	+	-	-	5:1--DDRF

M3 is closely linked to M1.

b. Using multiple SDRF from several loci

			F1—>												
Locus	P1	P2	1	2	3	4	5	6	7	8	9	10	11	12	Segregation
M1-1	+	-	+	+	+	+	+	+	-	-	-	-	-	-	1:1--SDRF
M1-2	+	-	+	+	-	-	-	-	+	+	+	+	-	-	1:1--SDRF
M4-1	+	-	+	+	+	+	+	+	-	-	-	-	-	-	1:1--SDRF
M4-2	+	-	+	+	-	-	-	-	+	+	+	+	-	-	1:1--SDRF

Since 1-1 and 4-1 link together and 1-2 and 4-2 link together, this is good evidence that these linkage groups are located on the same chromosome. (Only one of them, e.g., 1-1 and 1-2 also give this information, but we aren’t sure if they are duplications or not–i.e., whether the linkage groups associated with marker 1 represent the same locus or duplicated loci.)