Reconsidering gene identification using novel mapping strategies

Introduction

Once mutations have been identified in a screen, the next challenge is to determine which genes are affected by the mutation. We have already discussed the principle strategy to do that in the yeast part: Mapping. Mapping allows locating genes (mutations) along a chromosome. Here, we briefly recapitulate what we have already discussed about recombination mapping. Based on this, we will discuss recent approaches where instead of looking at visible phenotypes produced by mutations, mapping is done using molecular markers that do not produce a phenotype.

Complementation groups and mapping: a short reminder

We have also already discussed in the yeast part that before mapping the mutations, we use complementation tests to group mutations into complementation groups. A complementation group consists of mutations that do not complement each other, thus, these mutations are in the same gene. A complementation group is therefore equivalent to a gene. For mutations that are in the same complementation group we already know they are on the same location on a chromosome.

A complementation group contains all mutant alleles that were discovered in a screen. The alleles of a gene can affect the same phenotype, but they may differ in their strength and fall somewhere on a continuum, from moderate activity to very high activity. Geneticists refer to this as an allelic series. The members of this series can show various degrees of dominance to one another.

In cases where multiple alleles are present, dominance hierarchies can exist. Such a hierarchy, or allelic series as geneticists call it, is revealed by observing the phenotypes of each possible heterozygote offspring. To illustrate this, let's go back to the ABO system that determines the blood group in humans. The four blood types A, B, AB, and 0 are controlled by 3 alleles: I^A, I^B and i . i is recessive, and I^A and I^B are co-dominant, but both are dominant to i. Thus, the allelic series for the alleles determining the blood group is: $I^A = I^B > i$

For the alleles discovered in a screen that fall into one complementation group, allelic series can be determined to set up the hierarchy. This hierarchy of alleles can determine which allele from one complementation group is chosen for mapping: this is usually, the allele exhibiting the strongest phenotype that may be caused by a complete loss of function.

There are some general rules that apply to the strategy of mapping that we have already discussed, but will refresh here:

- Mapping is always a relative term. The position of gene can only be mapped relative to the position of a gene or marker for which the genomic location is already known (the reference points on a genetic map).
- Genes that are close together on the same chromosome are linked and do not segregate independently. They are inherited together.
- Linked genes lead to a larger number of progeny that resemble the parental class than expected if the two genes were to assort independently (i.e., unlinked genes),
- The underlying mechanism is recombination during meiosis and crossing over.
- The farther away genes are from each other, the greater is the chance that a crossover will occur between them.
- Recombination frequencies reflect the physical distance between genes (assuming that recombination is random).
- Recombination frequencies of two genes vary between 0 and 50%.

Linkage describes the fact that genes that are nearby on the same chromosome tend to stay together during the formation of gametes. Thus, they are inherited together. By the process of recombination (also called crossover), linkage can be abolished by separation of the genes and the exchange of genes between chromatids. Crossover can produce recombinant phenotypes, i.e., phenotypes that do not resemble the parental phenotypes.

Mapping by recombinant frequency

Genetic mapping is also known as gene mapping or chromosome mapping. Its purpose is to determine the linear order of linked genes along the same chromosome.

Genetic mapping experiments are typically accomplished by carrying out a testcross. This is a mating between an individual that is heterozygous for two or more genes and one that is homozygous recessive for the same genes. Genes located on different chromosomes show a recombination frequency of 50% (see figure 4-1A). Genes that are located far apart on the same chromosome show a recombination frequency of 50%, because crossover events are frequent enough that it is likely for recombination to occur (see figure 4-1B). The closer two genes are to each other on a chromosome, the smaller the recombination frequency will be (approaching 0% if the genes are very close). Thus, a recombination frequency of less than 50% indicates linkage.

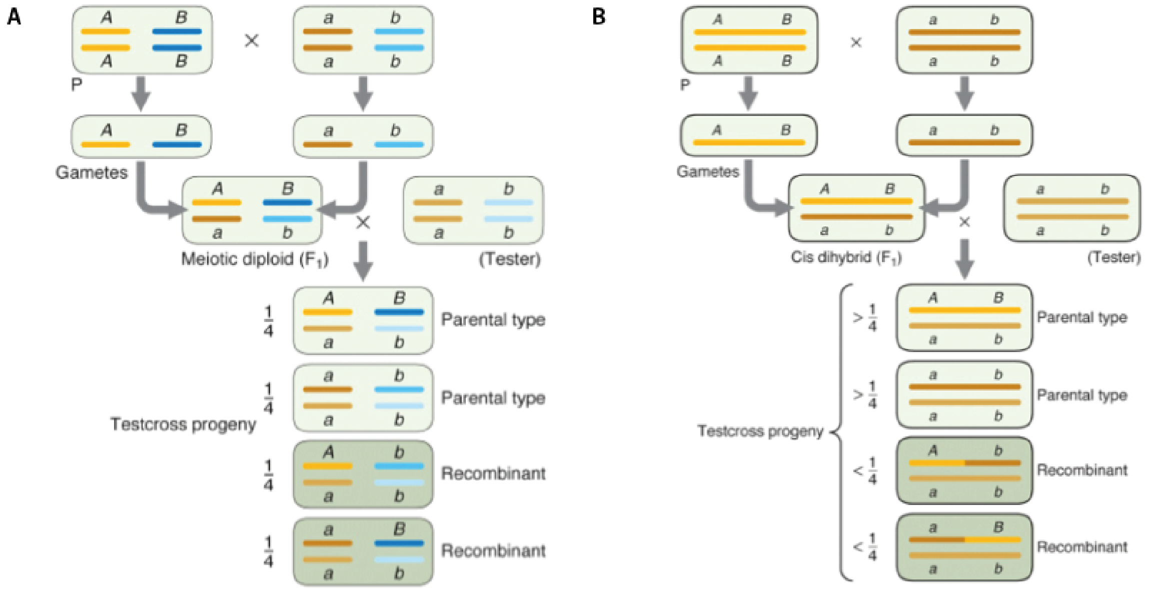

Figure 4-1 A testcross reveals the frequency of recombinants to determine genetic distances. (A) In a cross for two unlinked genes (located on different chromosomes), 50% of the progeny shows parental and 50% recombinant phenotypes. **(B)** In a cross for linked genes (located on the same chromosome), more than 50% of the progeny shows the parental phenotype and less than 50% the recombinant phenotype.

Thus, the recombination frequency is a measure for the distance between linked genes, and recombination frequencies can be used to establish linkage maps showing the relative position of several marker genes.

The first chromosomal linkage map

The fact that the frequency of crossover is a function of the genetic distance between genes was already realized in the early 19th century, when Thomas Morgan Hunt and Alfred Sturtevant performed genetic studies in *Drosophila*. By crossing different mutant flies with each other, they realized that they could use the frequency of recombination between the mutations to draw a chromosomal linkage map for the genes involved. They defined the unit of genetic distance as:

$$\rightarrow \text{Recombination frequency} = \frac{\text{Nr. of recombinants}}{\text{total number of flies}} \text{ x } 100 = \text{Map distance}$$

The unit of distance are called map units (m.u.) or centimorgans (cM).

Sturtevant hypothesized that calculating the recombination frequency could be used to produce a linear map that corresponds to the linearity of a chromosome. Indeed, the results from Sturtevant showed that the probability of crossover occurring between two genes that are relatively close on the same chromosome were additive. Thus, the map distances between genes are additive, which allows the construction of a linear chromosome map.

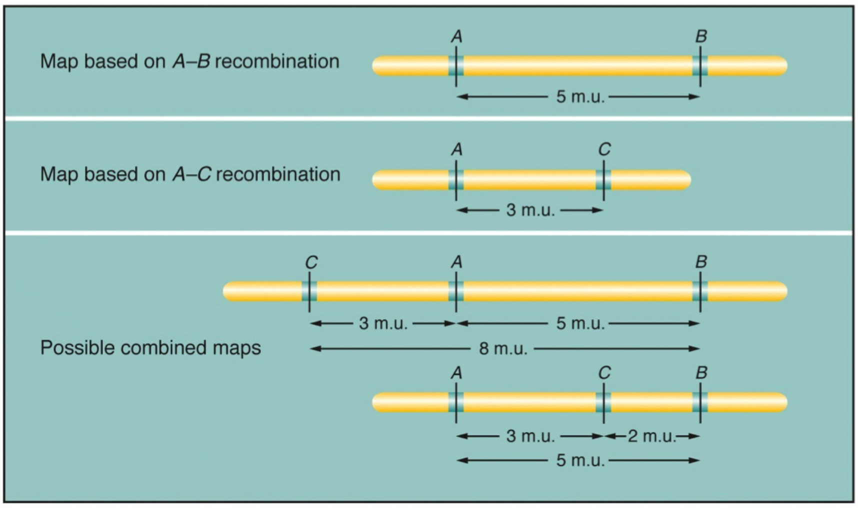

Figure 4-2 **Map distances are generally additive.** A chromosome region containing three linked genes is shown. Knowing the distances of A-B and A-C allows to place the genes A, B and C in two possible arrangements as shown in the lower panel.

Sturtevant and Morgan evaluated many different test crosses and finally published the first chromosomal map for mutations located on the X chromosome of *Drosophila*.

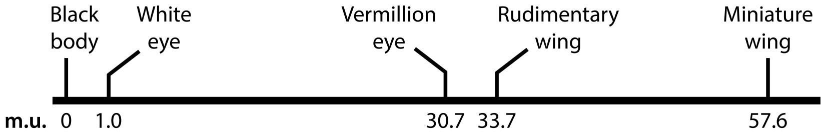

Figure 4-3 The first *Drosophila* linkage map as published by Sturtevant in 1913.

Mapping with molecular markers

So far, we have discussed in this course mapping genes using recombination frequencies by counting visible phenotypes that were produced by the different mutations involved. These phenotypes are often laborious to score and may interfere with the phenotype of the mutant of interest. Most importantly, because mutations with easily scored, viable phenotypes are relatively infrequent, the mapping resolution available using this approach is limited. A much higher degree of differences in the DNA between two chromosomes is available at the molecular level. These differences do not produce visibly different phenotypes, either because these differences are not located in genes or they are located in genes but do not alter the product protein. Such sequence differences can be thought of as molecular markers. Their position can be mapped by recombination frequencies in the same way as mutations producing visible phenotypes.

The two main types of molecular markers used in mapping are Single nucleotide polymorphisms (SNPs), small insertions and deletions (Indels), and simple-sequence-length polymorphisms (also called variable number tandem repeats, VNTR), repetitive DNA sequences of various lengths.

SNP and Indel polymorphisms are common in model organisms, such as *Drosophila* or *C. elegans*. The alleles of such markers are co-dominant and usually phenotypically neutral. Modern methods for scoring molecular variants offer the advantages of high throughput and automated scoring. The high density of SNPs in the genome and the availability of rapid scoring methods make for an attractive tool to analyze recombinant chromosomes.

A type of VNTR are copy number variations (CNV). A CNV is a DNA segment that is one kilobase (kb) or larger and that is present at a variable copy number in comparison with a reference genome. CNVs are relatively frequent (e.g., between 5-9% of the human genome are estimated to be CNVs). While some of them have no apparent influence on the phenotype, while others have been definitively linked with a disease. Perhaps the best-defined and most widely known CNVs are the trinucleotide repeats (TNRs), which consist of three nucleotides repeating in tandem. CNVs resulting in a massive increase of TNR copy have been shown to be a cause of Huntington's disease, with the number of repeats varying in both normal and affected individuals.

You can find additional information about different types of molecular markers here:
https://www.ncbi.nlm.nih.gov/books/NBK21116/

The general mapping strategy using molecular markers

We will now explain the general principle of mapping a mutation in the genome using a genetic marker and, for simplicity, will only consider a single chromosome. To further simplify things, we first look at a haploid organism, such as the fungus *Neurospora crassa*.

Two strains of *Neurospora crassa*, Mauriceville and Oakridge, have five different genetic markers, with known location, that each has two different alleles, *M* (for the Mauriceville strain) and *O* (for Oakridge, see figure 4-4). In Mauriceville, all markers have the *M* allele, and in Oakridge all markers have the *O* allele. While Mauriceville shows the wild-type phenotype (+), Oakridge has been mutagenized and now contains the mutation of interest, *m*, whose location is unknown.

The mutant strain Oakridge is now crossed with the wild-type strain Mauriceville and the progeny analyzed. For every marker, the allele present is compared to the resulting phenotype in order to identify the marker location where the marker allele correlates strongest with the mutant phenotype. In our example, marker 4 is always present as the *O* allele in the progeny showing the mutant phenotype, and always present as the *M* allele in the progeny showing the wild-type phenotype. Since the allele present at this marker position correlates strongest (i.e., shows the strongest linkage) with the resulting phenotype, this marker must be closest to the mutation *m* and thus allows pinpointing the mutation to this location in the genome.

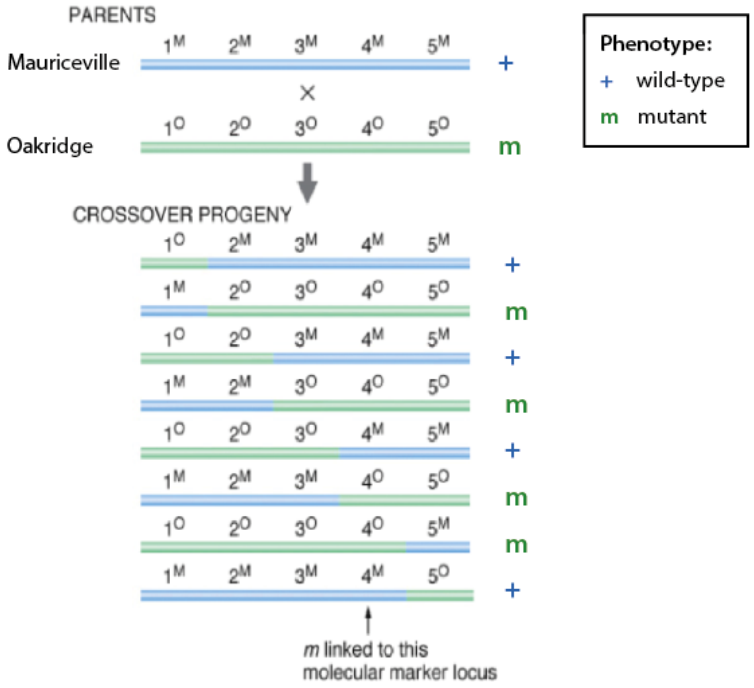

Figure 4-4 Principle of marker mapping. Two strains of the fungus *Neurospora crassa*, Mauriceville and Oakridge, are represented in blue and green, respectively. Both strains contain five different genetic markers (1-5) with known location, with each having two different alleles *M* or *O*. Mauriceville is a wild-type strain (+), while Oakridge shows a mutant phenotype (m). Mauriceville and Oakridge are crossed and the progeny analyzed for the occurrence of the alleles markers in respect to the resulting phenotypes (+ or m) in order to map the mutation *m* to one of the molecular markers.

Principle of SNP mapping

SNPs (Single-Nucleotide Polymorphisms) are DNA polymorphisms at a specific position in the genome that differ in only a single nucleotide, with each variation of the nucleotide being present to some degree within a population (see figure 4-5). SNPs are thus well suited to be used as molecular markers for gene mapping. They are frequently found in the genome and thereby allow for gene mapping to a few kilobases, i.e., at a very high resolution. For example, in the human population there exists about one SNP every 800 base pair, though the distribution of the SNP (SNP density) is inhomogeneous and SNP frequencies vary from population to population.

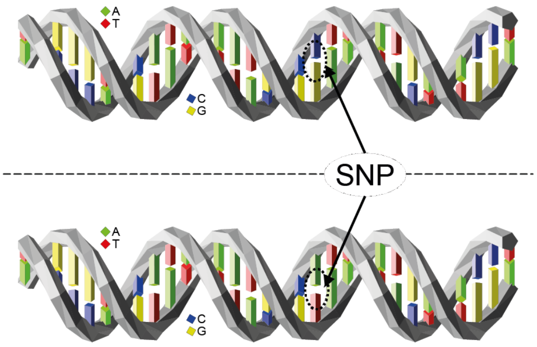

Figure 4-5 Single Nucleotide Polimorphism (SNP). A SNP is a change of a nucleotide at at single base-pair location on DNA, that is present with a certain frequency in the population. (adapted from wikipedia: Single Nucleotide Polymorphism)

In order to map a mutation of interest to a SNP, the SNP has to be detected and identified molecularly. Most popular, especially in human genetics, is the use of so-called SNP microarrays. SNP microarrays contain many different nucleotide sequences on a chip allowing to simultaneously bind and identify thousands of SNPs (up to 500'000) over the entire genome (see figure 4-6).

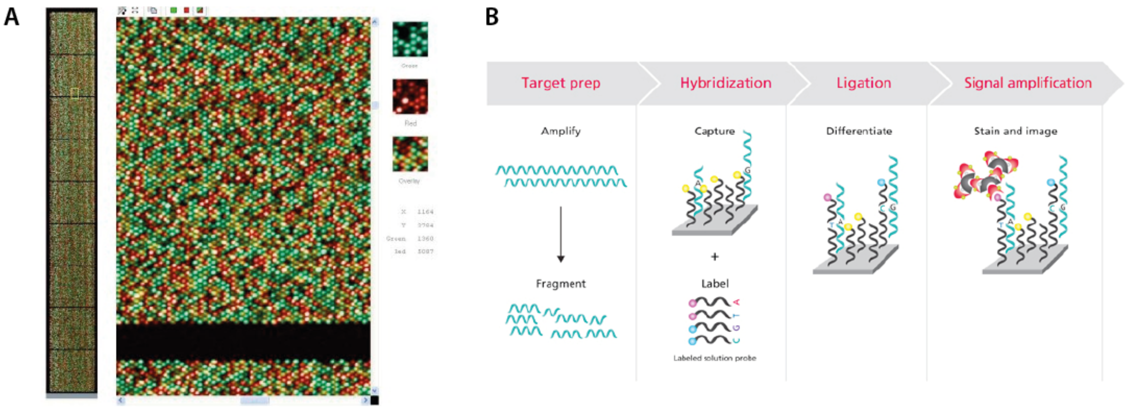

Figure 4-6 **(A)** Example of a SNP microarray: The Affymetrix SNP array 6.0 contains probes for the different alleles of 1.8 million SNPs and CNVs (copy-number variation = repeated regions in the genome with variable copy number) molecular markers. **(B)** The genomic sequence to be analyzed is PCR amplified and randomly fragmented. The fragments are hybridized to a SNP microarray and labeled with a fluorescent dye. By analyzing the resulting fluorescence, the SNP can be identified.

Example of a SNP mapping approach in *C. elegans*

The most commonly used mapping strategy in *C. elegans* employs single nucleotide polymorphism (SNP)-based mapping. In this strategy, DNA sequence polymorphisms between the wild-type *C. elegans* strain (Bristol) and a closely related strain (Hawaiian) are used as genetic markers. Similar to humans, these two strains differ by around one SNP or Indel every 800 base pairs. These differences are broadly dispersed throughout the genome, such that every gene in the genome is marked by multiple SNPs or Indels. The two strains are crossed and the mutant F2 progeny of such a cross are analyzed for their distribution of SNP markers.

SNP mapping is usually done in two phases. The first phase, chromosome mapping, is similar to traditional two-factor mapping and seeks to identify the relevant chromosome and rough position of the gene of interest. The second phase, interval mapping, seeks to place the gene of interest in an interval between two SNPs, and can be used iteratively to fine map the gene.

In the example shown in figure 4-7, the Bristol strain used as one of the parental strains contains the recessive mutation to be mapped and a visible, recessive marker as reference to identify recombinants on the chromosome of interest. Crossing this Bristol to a Hawaiian strains leads to a heterozygous F1 generation. In these animals, recombination can take place, leading to an F2 progeny that can be screened by looking at the visual marker. In this F2 progeny, recombination leads to animals containing the mutation, but that do not show the marker (because they are heterozygous for the marker, left situation in figure 4-7). On the other hand, there are animals that show the marker, but not the mutation (right situation in figure 4-7).

Genomic regions close to the mutation of interest show a small incidence of Hawaiian SNPs, because no recombination has taken place at this position. Thus, in the mutant animals (left, figure 4-7), the piece of Bristol sequence that is present in every mutant animal must be close to the mutation. Thus, this region must be "free" of Hawaiian SNPs. On the other hand, unlinked regions (regions that are far away from the mutation) contain an even distribution of Hawaiian and Bristol SNPs.

Subsequent genotyping of SNPs by PCR of single recombinants across the region reveals which of the SNPs are always Bristol. These must be linked to the mutation and thus gives the genetic position of the mutation. The big advantage is that multiple markers can be followed in a single cross by carrying out PCRs on the DNA samples of the progeny.

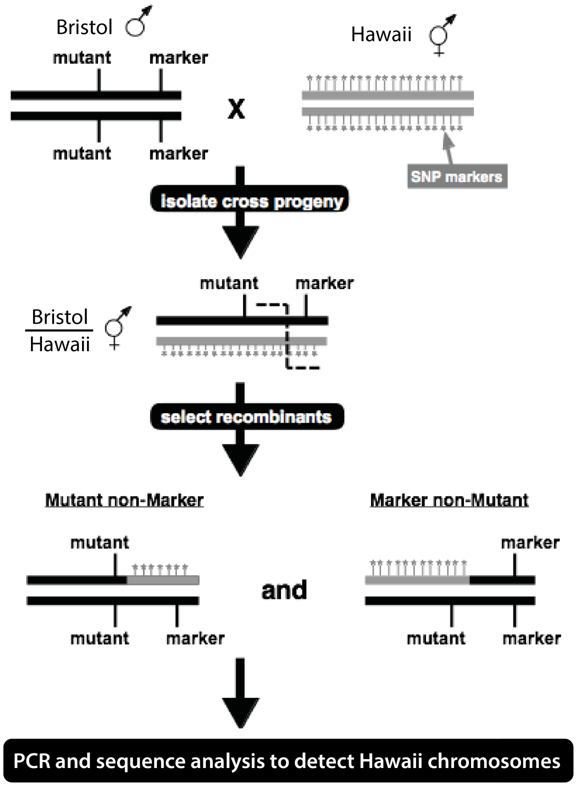

Figure 4-7 **Strategy for high-resolution SNP mapping in *C. elegans*.** Because of geographical separation, several million years of evolutionary drift have led to a considerable number of genetic differences (DNA polymorphisms) between the Hawaiian and Bristol *C. elegans* populations. The high density of SNPs in the genome allows to fine map mutations even to a single gene.

Combining whole-genome sequencing and SNP mapping

SNP mapping has been used since 2001 to map mutations. An approach that went one step further incorporated SNP mapping into whole-genome sequencing (WGS) to identify *C. elegans* mutants. A very similar protocol is nowadays also used in plants and *Drosophila*. The idea was to first map the mutation by SNP mapping as described above; however, instead of analyzing each individual animal of the F2 generation for the SNP allele to be present, the F2 animals carrying the mutant phenotype were singled out to fresh plates. These F2 animals were left to produce progeny by self-fertilization, and this F3 progeny of many (around 200) F2 recombinants were pooled, DNA was prepared and the pool was subjected to whole genome sequencing (WGS). This mapping strategy is called "bulk segregant analysis".

In regions unlinked to the mutation (far away), the parental chromosomes will recombine in a largely non-biased manner. So as long as enough recombinants are pooled, unlinked loci will appear in even ratio of Hawaiian vs. Bristol SNPs. Thus, the sequencing will show a 50/50 ratio of Hawaiian versus Bristol SNPs (this is indicated by the "pileup" window in figure 4-8). However, the closer a SNP is to the mutation, the more rare it is to find a recombination event between this SNP and the mutation.

In this case, the sequencing will detect less Hawaiian SNPs than Bristol SNPs in regions that are closer to the mutation. In the areas very near to the mutation, the sequencing will detect only Bristol SNPs, since the SNP and the mutation are too close for recombination to take place.

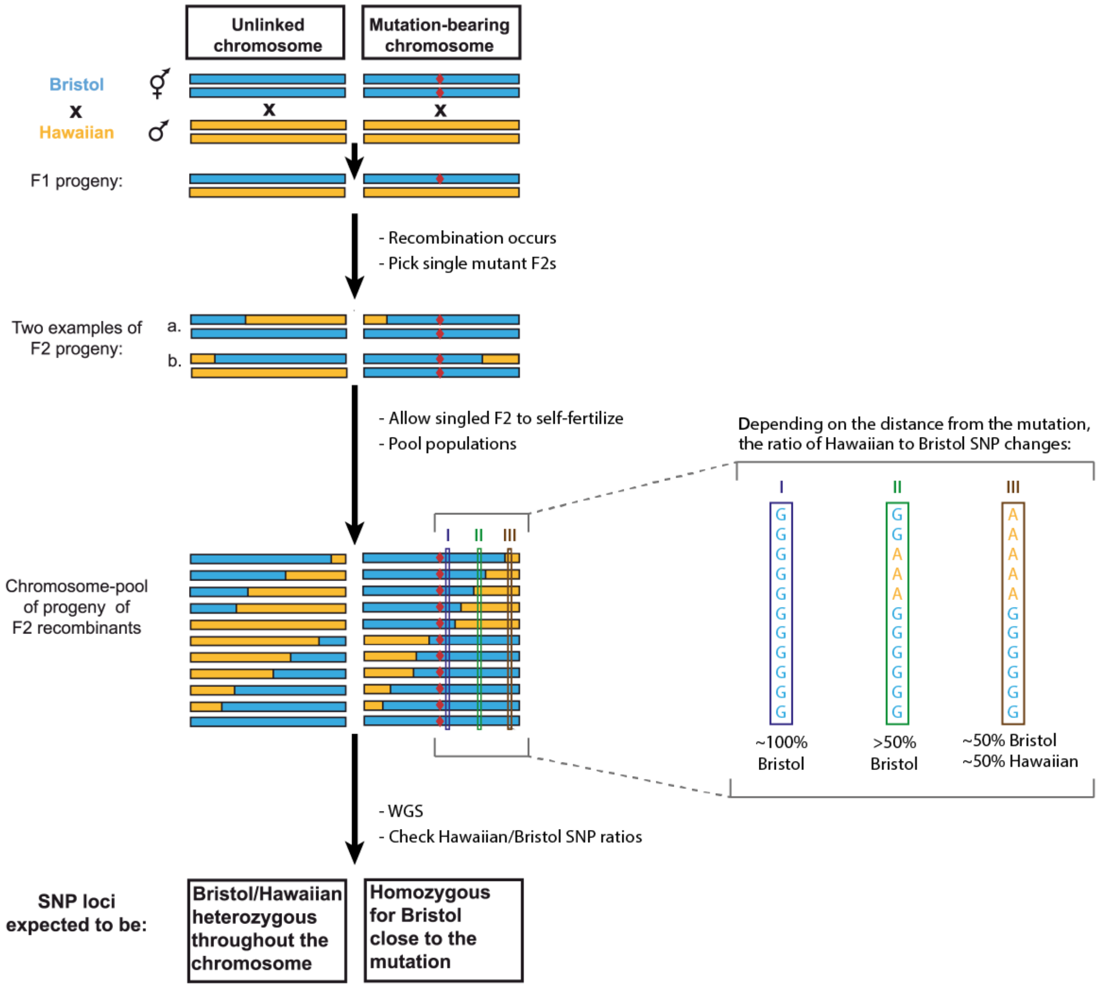
Unlinked chromosome
Mutation-bearing chromosome
Bristol ⚥
x
Hawaiian ♂
x
x
F1 progeny:
- Recombination occurs
- Pick single mutant F2s
Two examples of F2 progeny:
a.
b.
- Allow singled F2 to self-fertilize
- Pool populations
Chromosome-pool of progeny of F2 recombinants
I II III
Depending on the distance from the mutation, the ratio of Hawaiian to Bristol SNP changes:
I
G G G G G G G G G G G
~100% Bristol
II
G G A A A G G G G G G
>50% Bristol
III
A A A A A G G G G G G
~50% Bristol ~50% Hawaiian
- WGS
- Check Hawaiian/Bristol SNP ratios
SNP loci expected to be:
Bristol/Hawaiian heterozygous throughout the chromosome
Homozygous for Bristol close to the mutation

Figure 4-8 Principle of the WGS-SNP strategy. The mutation of interest is indicated by the red diamond. SNP mapping is done based on differences in the Bristol and Hawaiian strains as described before, with the difference that the single F2 animals were not analyzed for the presence of SNPs, but were left to produce F3 progeny by self fertilization. The pooled F3 is sequenced and the ratios of Hawaiian vs. Bristol SNPs are analyzed by piling up the number of reads that are mapped to a specific position in the genome (here, an A in the Hawaiian and a G in the Bristol strain). The ratios of Hawaiian/Bristol SNPs reflect the relative distribution of recombinants within the pooled F3 progeny. (Doitsidou et al., *PLoS ONE*)

Integrated genome maps

Until now, we have concentrated on the generation of genetic maps that rely on information gained from recombination. They show the loci of genes for which mutant alleles (and their mutant phenotype) have been found. The position of these loci is determined on the basis of recombination frequencies. We have also seen how sites of molecular diversity (that are not associated with mutant phenotypes) can be incorporated into recombination maps (see SNP mapping above). Also these molecular makers are mapped by recombination and then used to navigate toward a gene of interest. A recombination map thus represents the arrangement of genes on a chromosome and distances are indicated in centiMorgans. However, these maps are hypothetical constructs.

Physical genome maps are constructed by identifying the base-pair sequence of genes as well as their position along the chromosome, and thus provide exact DNA sequence information. The units of distance on a physical map are numbers of DNA bases (or kilobases).

Whole genome sequencing provides the most thorough tool to construct a complete physical genome map at single base-pair resolution. Therefore, large numbers of small genomic fragments are sequenced and then assembled into a whole sequence. This sequence is then scanned by a computer, looking for gene-like sequences to assign genes to the sequence, a process called annotation We will discuss this process in more detail later in this course.

Physical maps are now available for most of the genetic model organisms. However, recombination maps still provide valuable information, and integrated genome maps combine the information of both types of maps to show entire genomes of different strains or organisms that display their coding sequences, their polymorphisms, genetic correlations, and functional elements.

This principle is illustrated in figure 4-9, which shows a physical and a recombination map of the same region of the genome. For some of the genes, e.g., A and B in this example, a function may have been discovered. For other genes, e.g., the one in the middle in figure 4-9, a certain phenotype has been discovered by a mutation in a screen. To determine its function, a physical map is helpful, and the genes present in this region may all represent candidate genes. Any of those could be the gene of interest. If for one of these genes, a function is known in another organism, this suggests that this gene may have a similar function in the organisms that is currently under investigation. In this way, the phenotype mapped on the recombination map can be linked to a function deduced from the physical map. Thus, both maps contain information that is complementary: the physical map shows a gene's possible action on the cellular level while the recombination map reveals the effect of the gene on the phenotypic level.

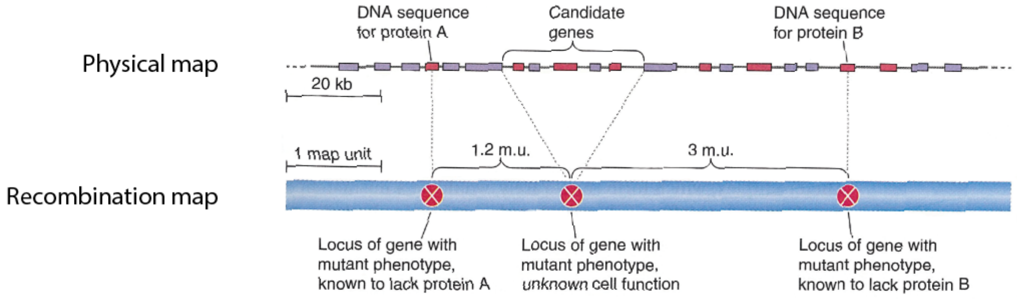
DNA sequence for protein A
Candidate genes
DNA sequence for protein B
Physical map
20 kb
1 map unit
1.2 m.u.
3 m.u.
Recombination map
Locus of gene with mutant phenotype, known to lack protein A
Locus of gene with mutant phenotype, unknown cell function
Locus of gene with mutant phenotype, known to lack protein B

Figure 4-9 The alignment of physical and recombination maps allow to connect a phenotype with the function of a gene that is currently unknown (purple), or suspected (red) from another organism.

Integrated genome maps are particularly helpful in the analysis and research of diseases. By analyzing all genomic and genetic information available about a disease, integrated genome maps are crucial for determining the genetic origin and reason for a disease, for identifying surrogate genetic markers for the disease, as well as for finding correlations between diseases and between disease and genotype.

Integrated genetic maps for any organisms are nowadays easily accessible throughout the internet, for example:

- Human: www.ensembl.org
- *C. elegans*: www.wormbase.org
- *D. melanogaster*: www.flybase.org