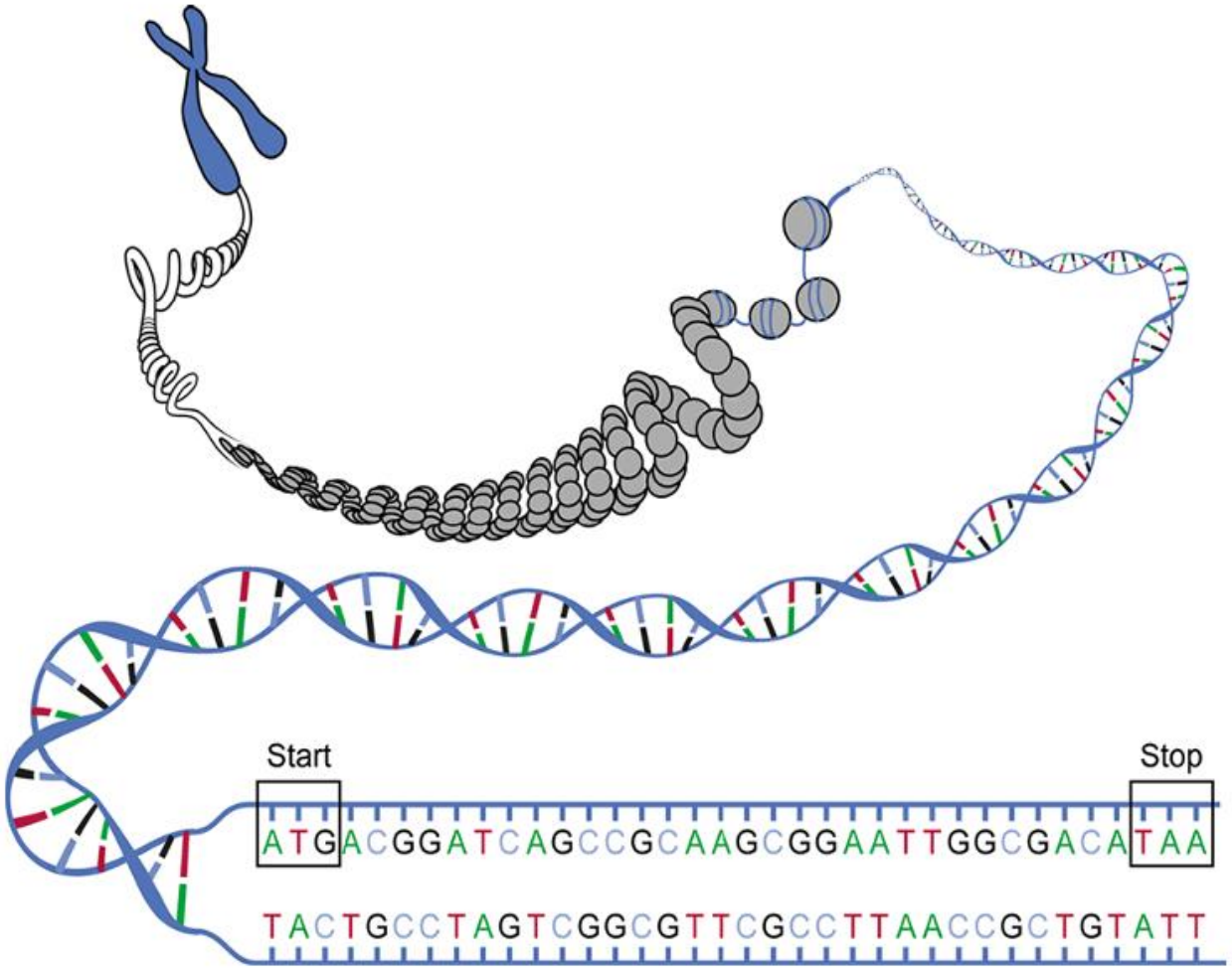
INSTITUTE OF PLANT SCIENCES

Sant'Anna
School of Advanced Studies – Pisa

Advanced Genomics

Genetic maps

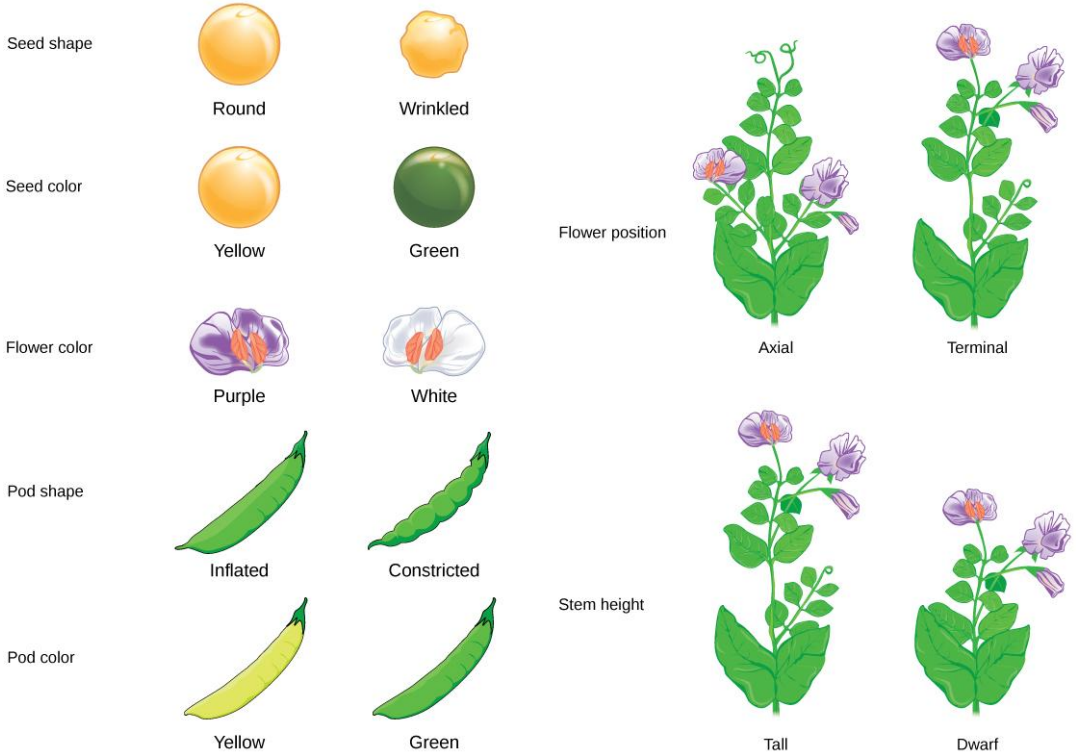


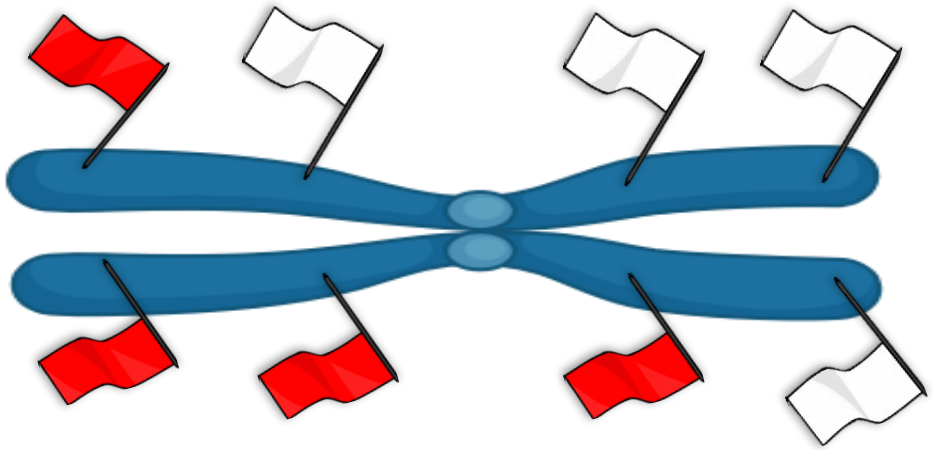
We now understand variation, and the many different elements making up genomes

If we want to be in the position to read the DNA sequence from start to end of any given chromatid, we need to understand how to pinpoint the location of individual loci

The key to reconstruct the linear organization of genomic loci: inheritance

Mendel used morphological traits to understand inheritance and, by extension, genetics; we now (sort of) understand genetics and use molecular markers to follow inheritance





Mendel laws

1

Dominance

- If the two alleles of an inherited pair differ, then one determines the organism's appearance and is called the dominant allele; the other has no noticeable effect on the organism's appearance and is called the recessive allele

2

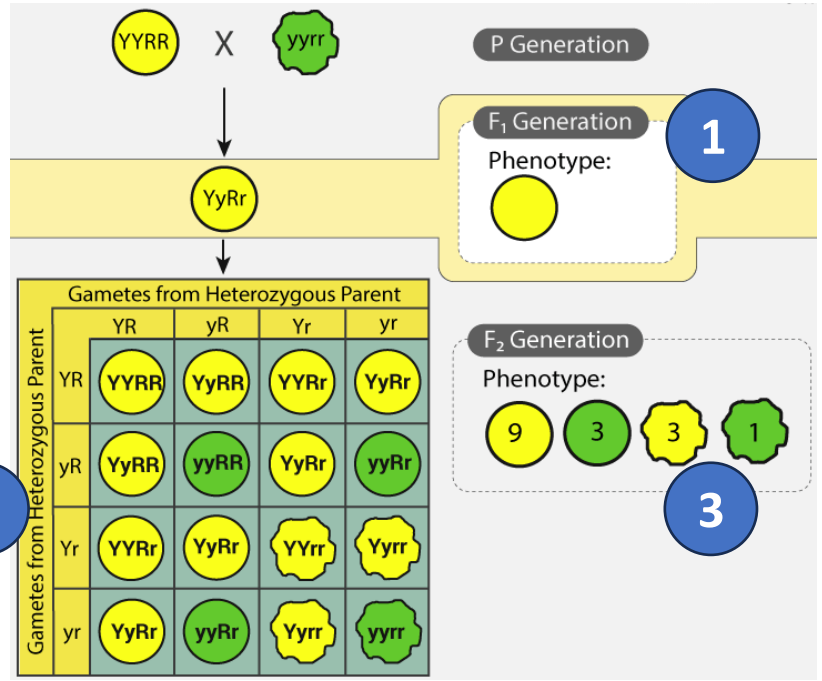
Segregation

- Every individual organism contains two alleles for each trait, and alleles segregate during meiosis such that each gamete contains only one of the alleles

3

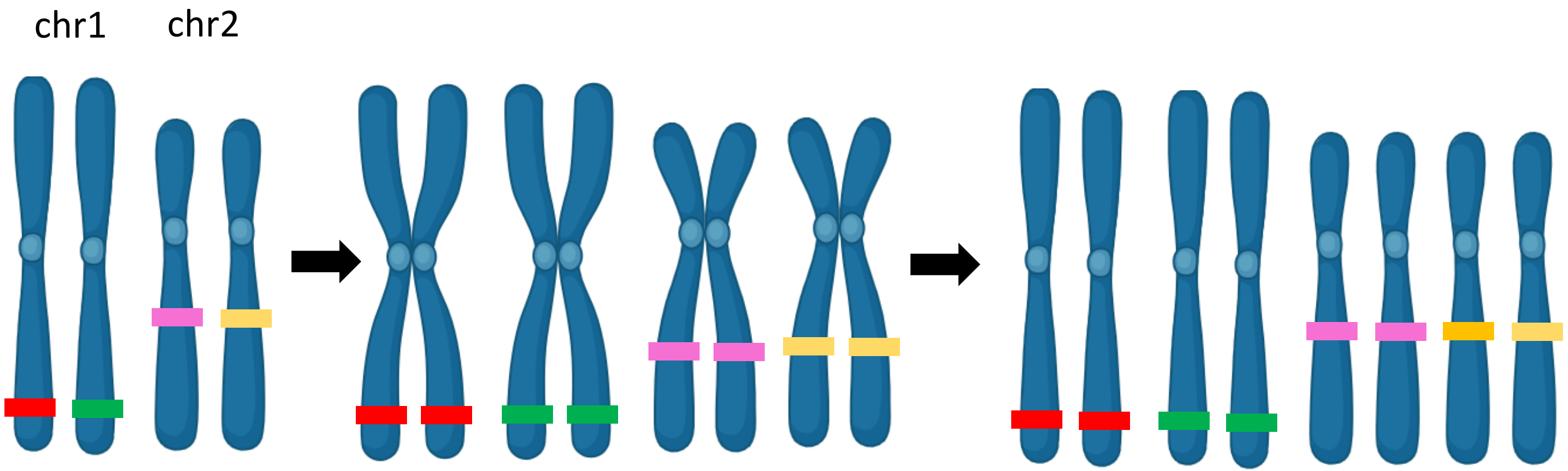
Independent Assortment

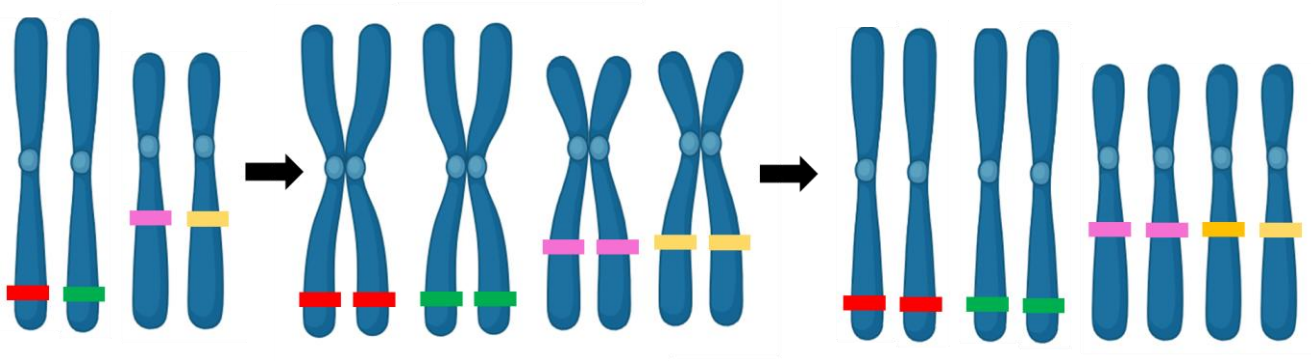
- Unlinked or distantly linked segregating genes pairs behave independently

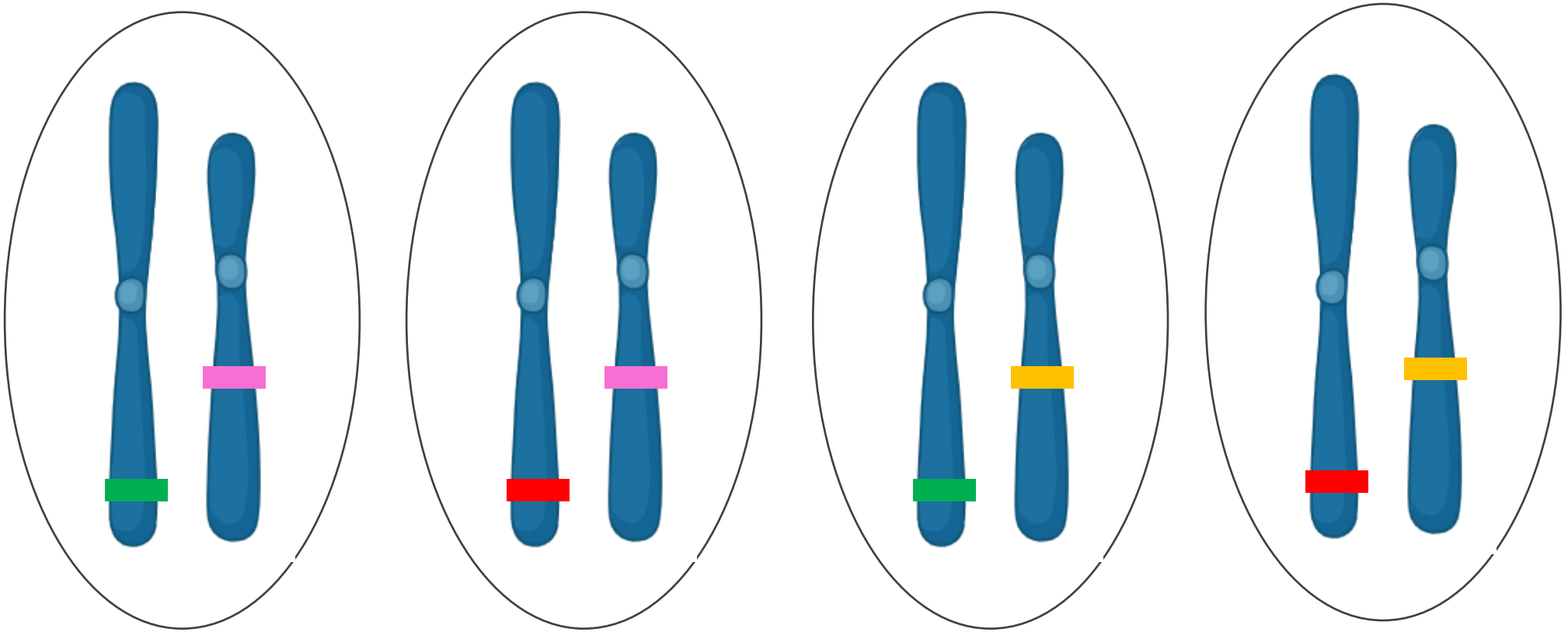


Segregation of unlinked loci/genes

- Assume diploid individuals; loci segregate during meiosis
- Unlinked loci (i.e. on different chromosomes) segregate independently (these are the mendelian factors!)



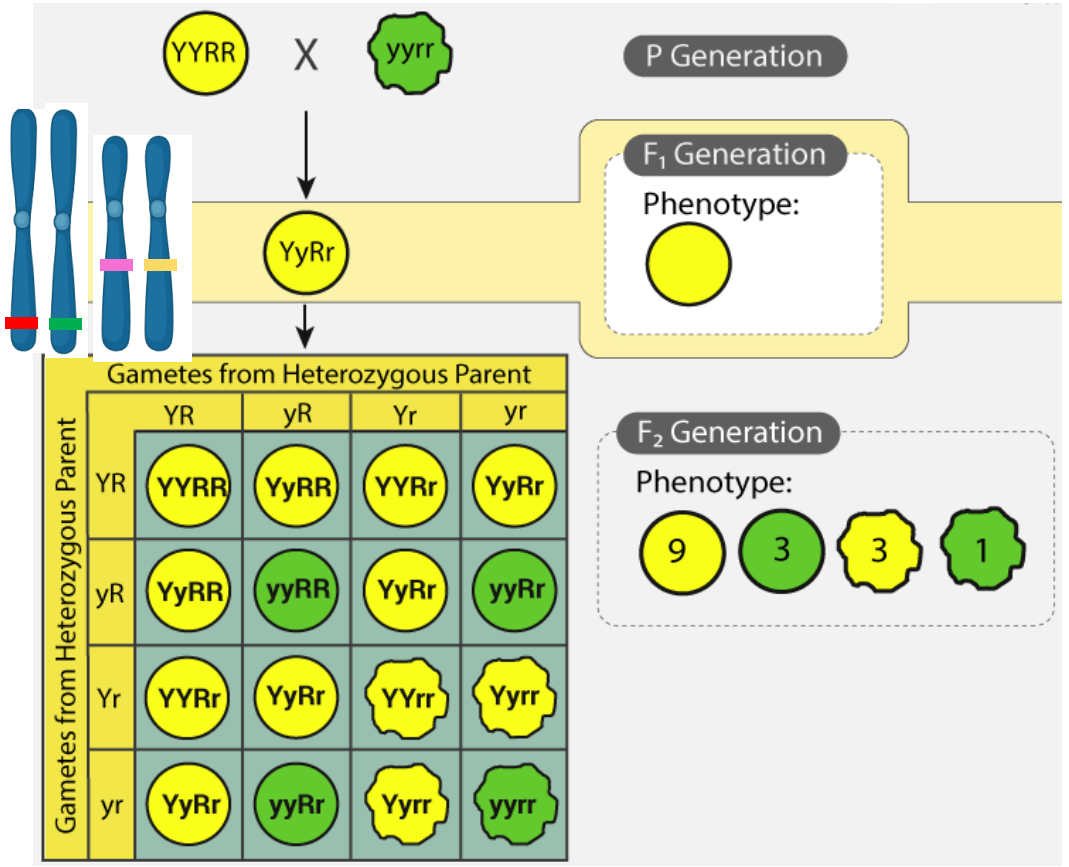




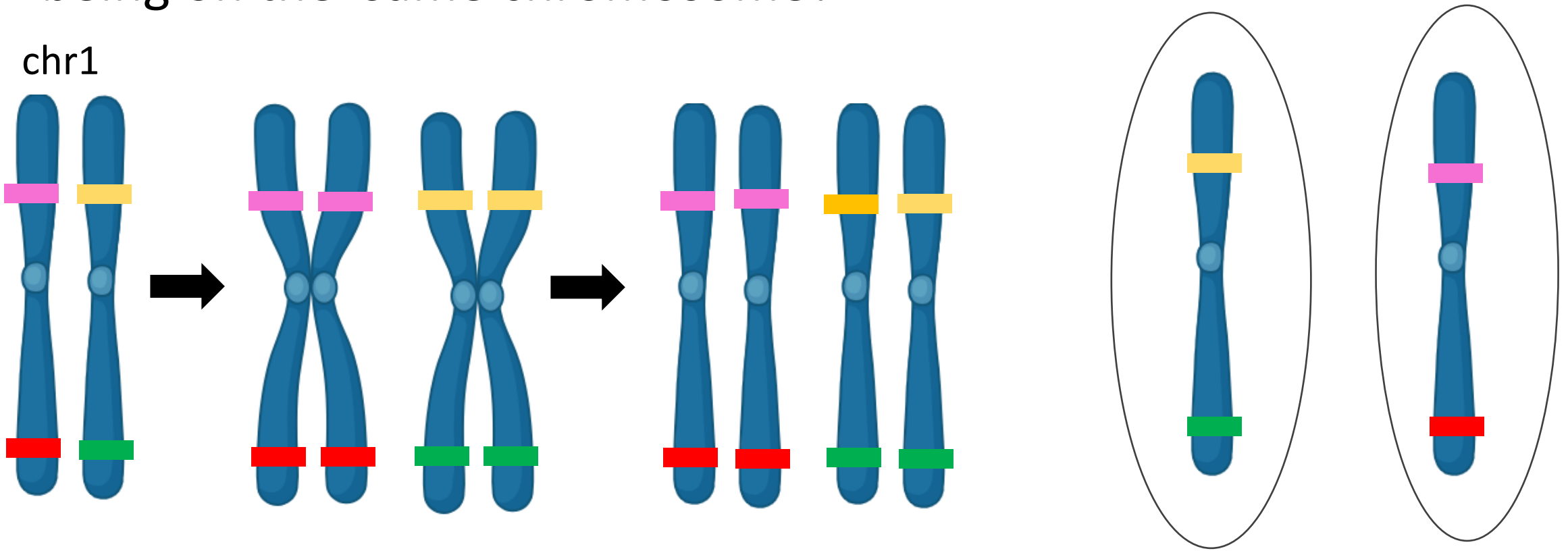
Gametes contain all possibile allelic combinations

Unlinked loci/genes assort independently

- When loci are on different chromosomes, they travel independently during meiosis → segregation leads to independent assortment
- This results in four possible gametes (as long as we are looking at two genes)
- This is where the dihybrid cross ratio 9:3:3:1 recombination ratio comes from

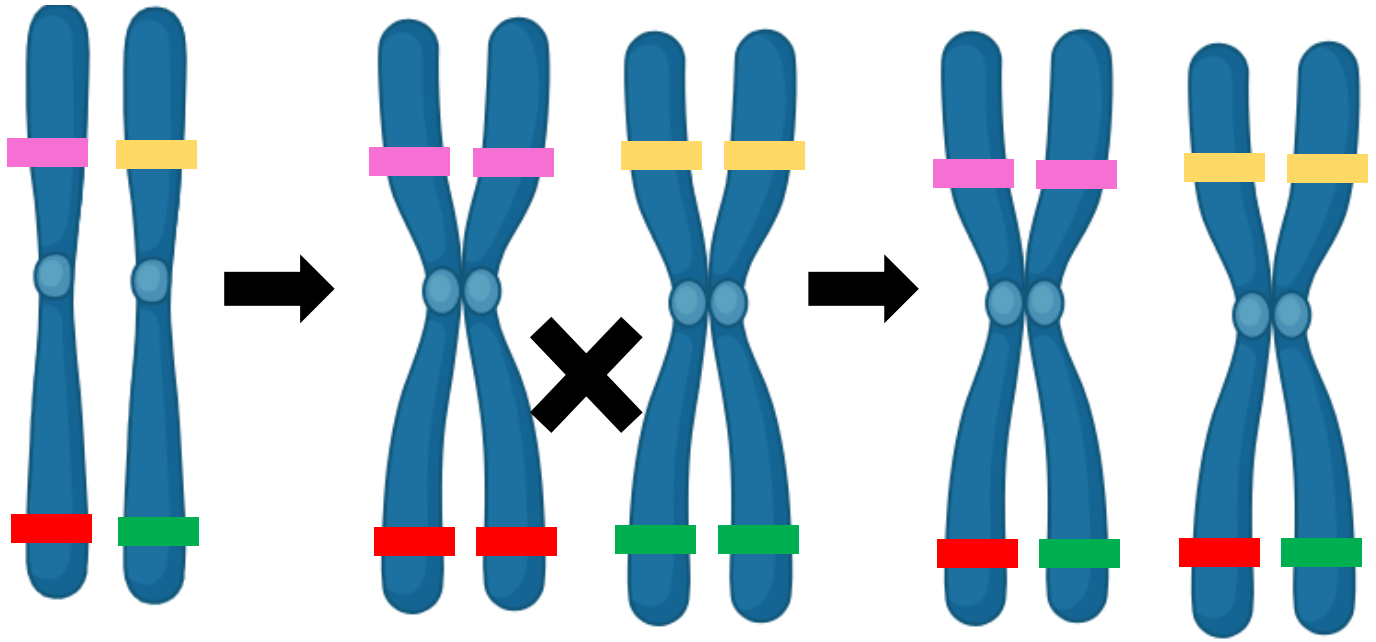


What if now we consider the same genes & same alleles being on the same chromosome?

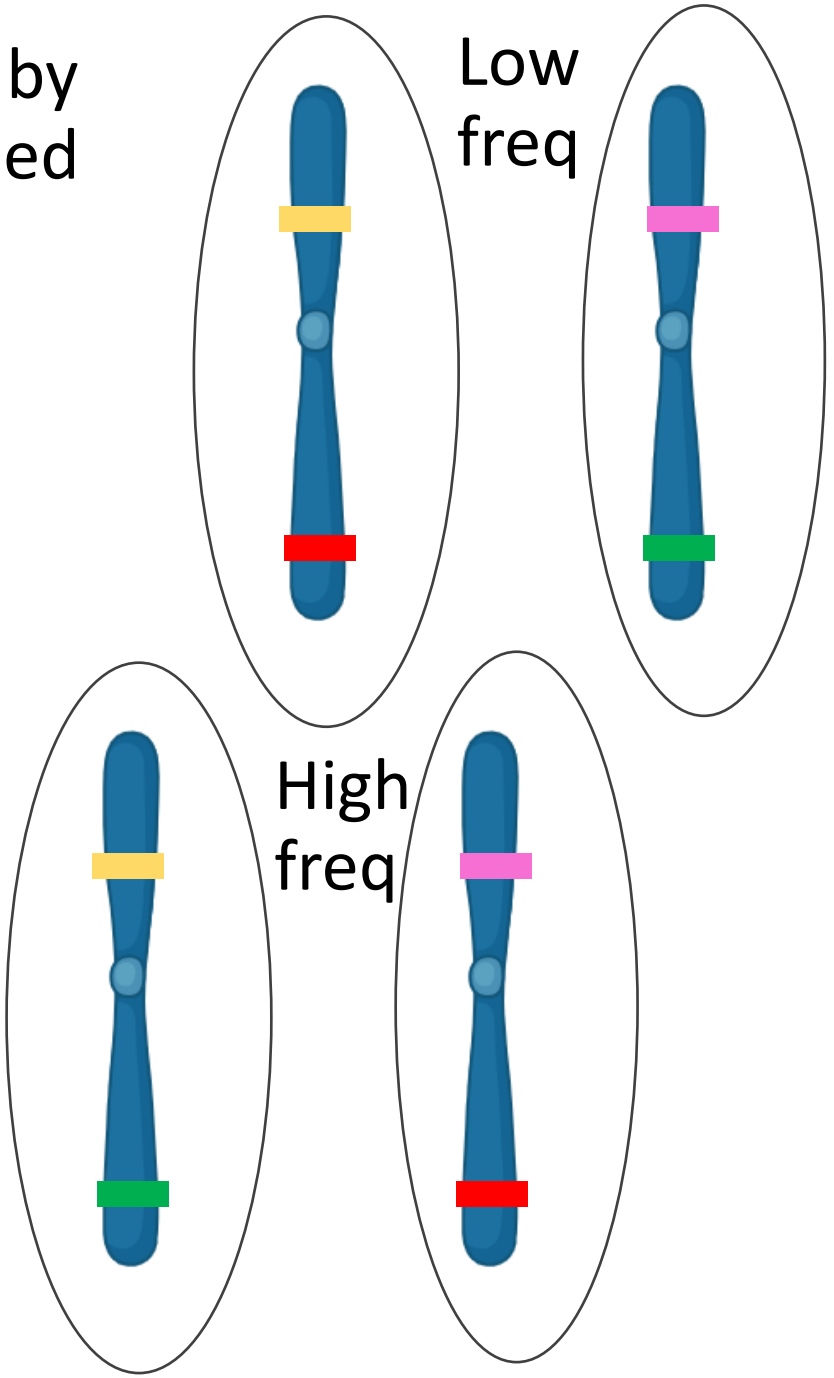


If there is no exchange of genetic materials b/w chromatids, then the alleles are inherited in the same pairing as in the parental lines

Crossing over AKA recombination is the mechanisms by which alleles on the same chromosome can be shuffled

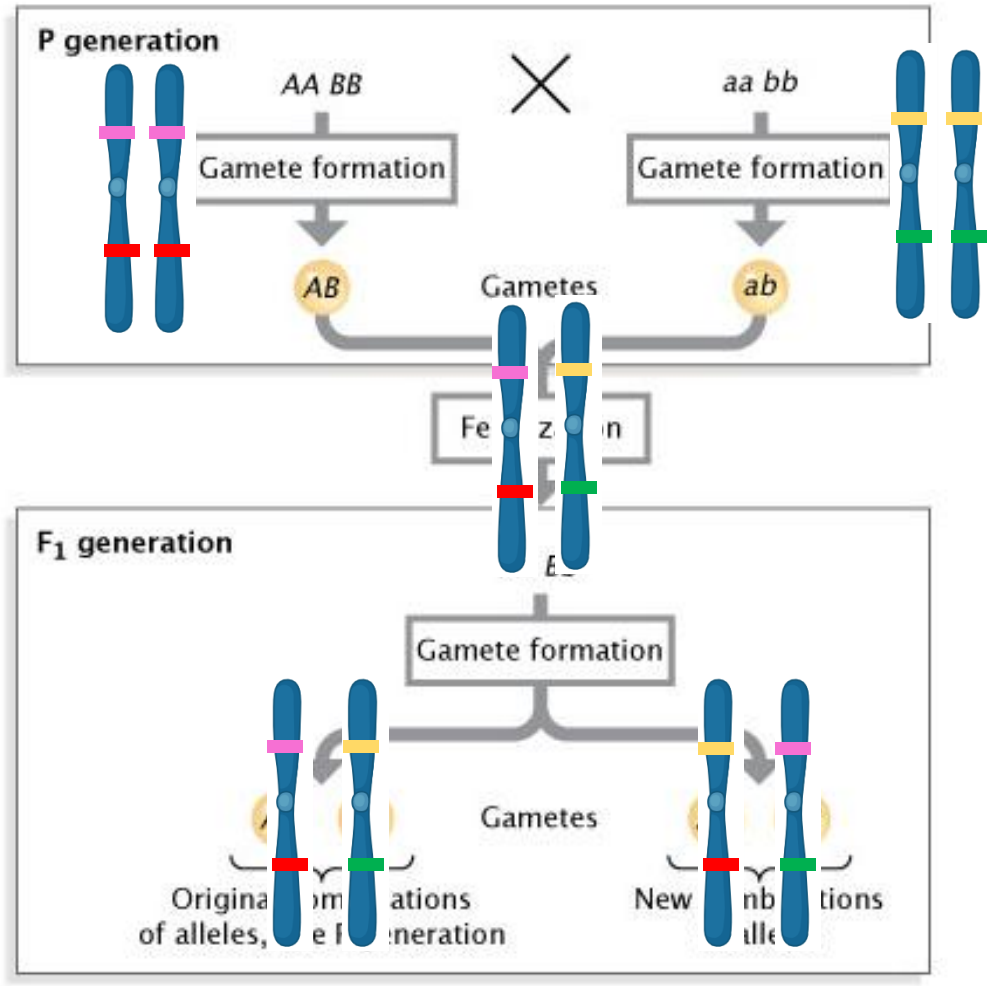


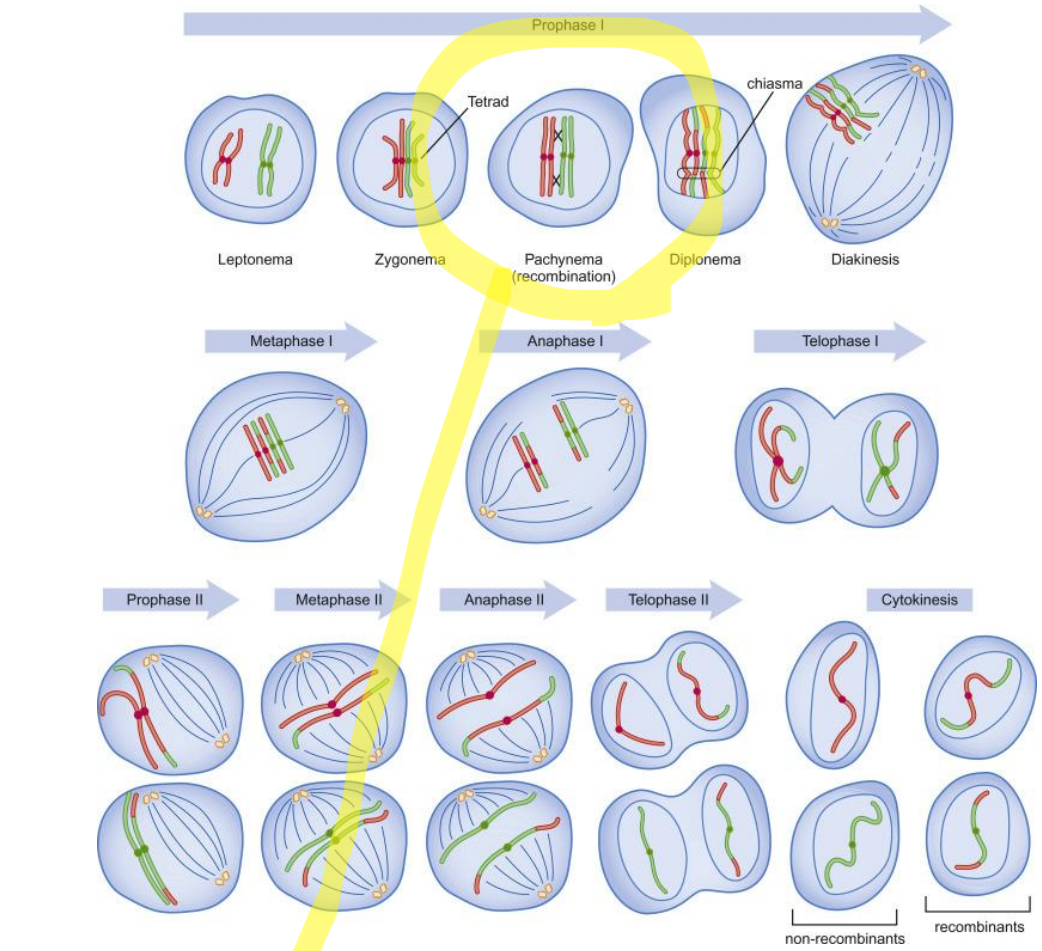
Looking at how often alleles are shuffled we can understand with which frequency they recombine; hence their distance



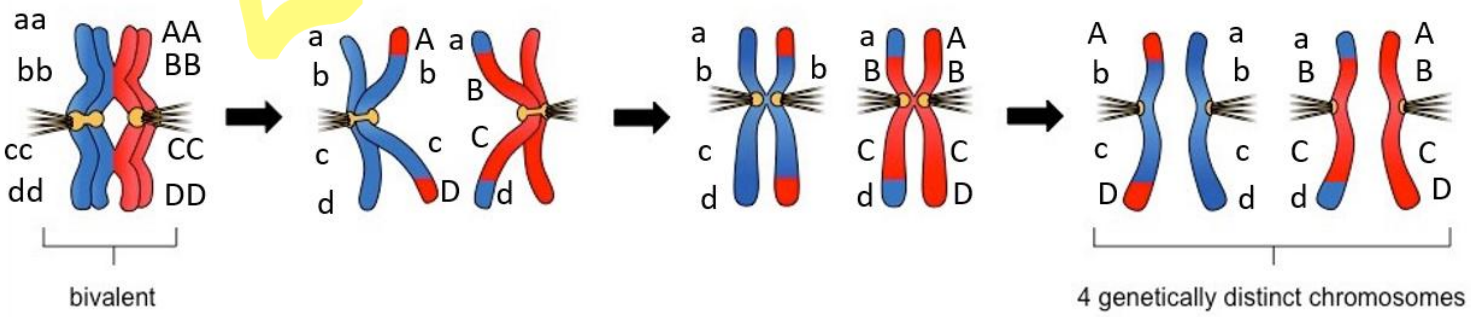
Linked loci/genes **don't** assort independently

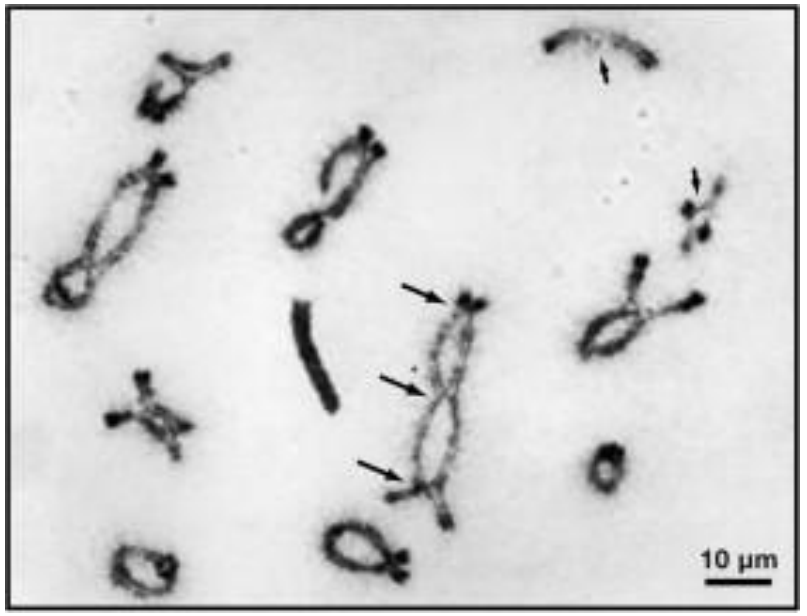
- When loci are on the same chromosome (AKA linkage group), they are connected and can't assort independently
- They travel together landing in the same gamete. A package deal, two for one
- UNLESS, recombination occurs and mixes up alleles
- This still results in four possible gametes (as long as we are looking at two genes), but frequencies of each combination differ according to the frequency of crossing over

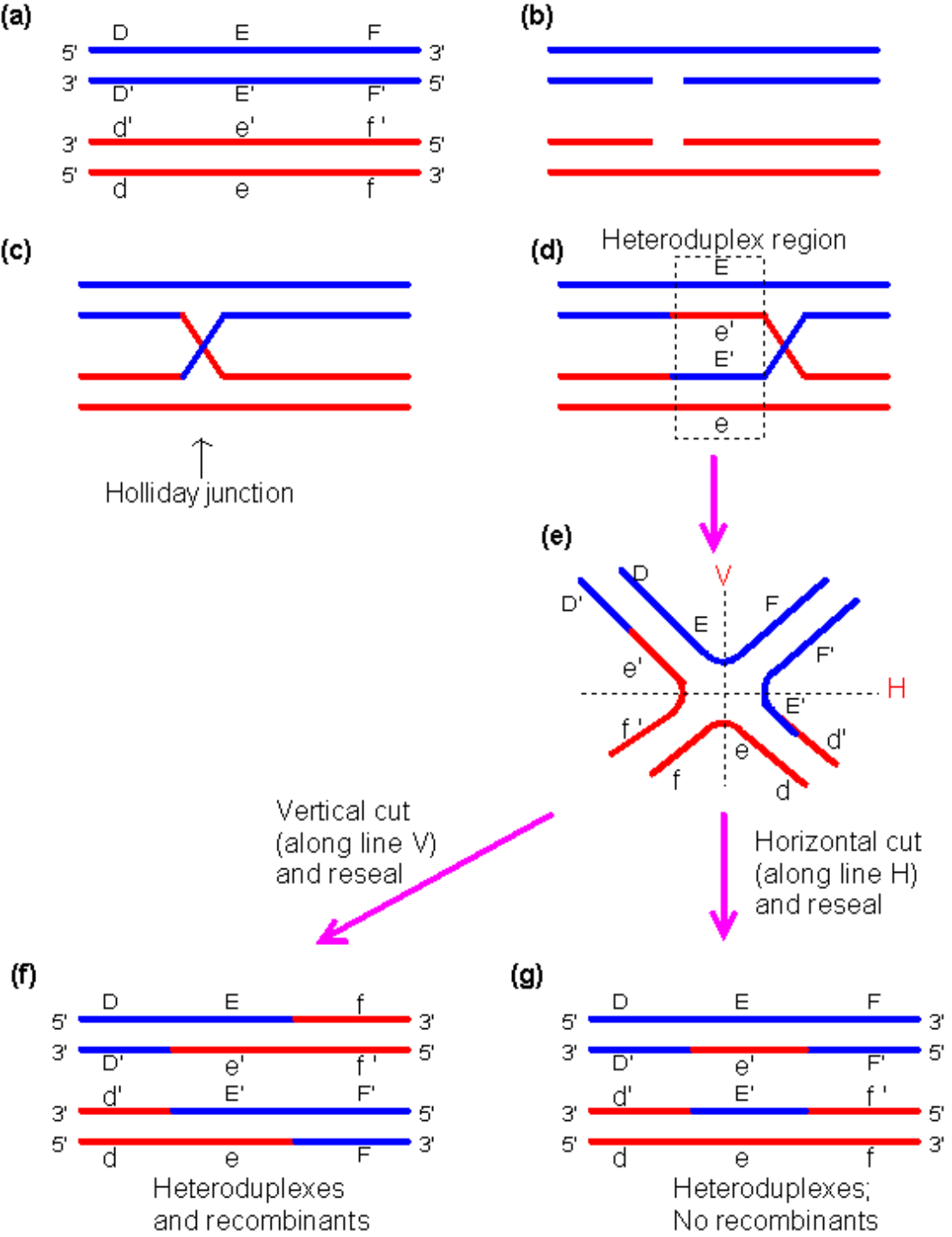




- Crossing over is an exchange of genetic material occurring between homeologous chromosomes (based on sequence homology and position)
- New allelic combinations are formed at each generation





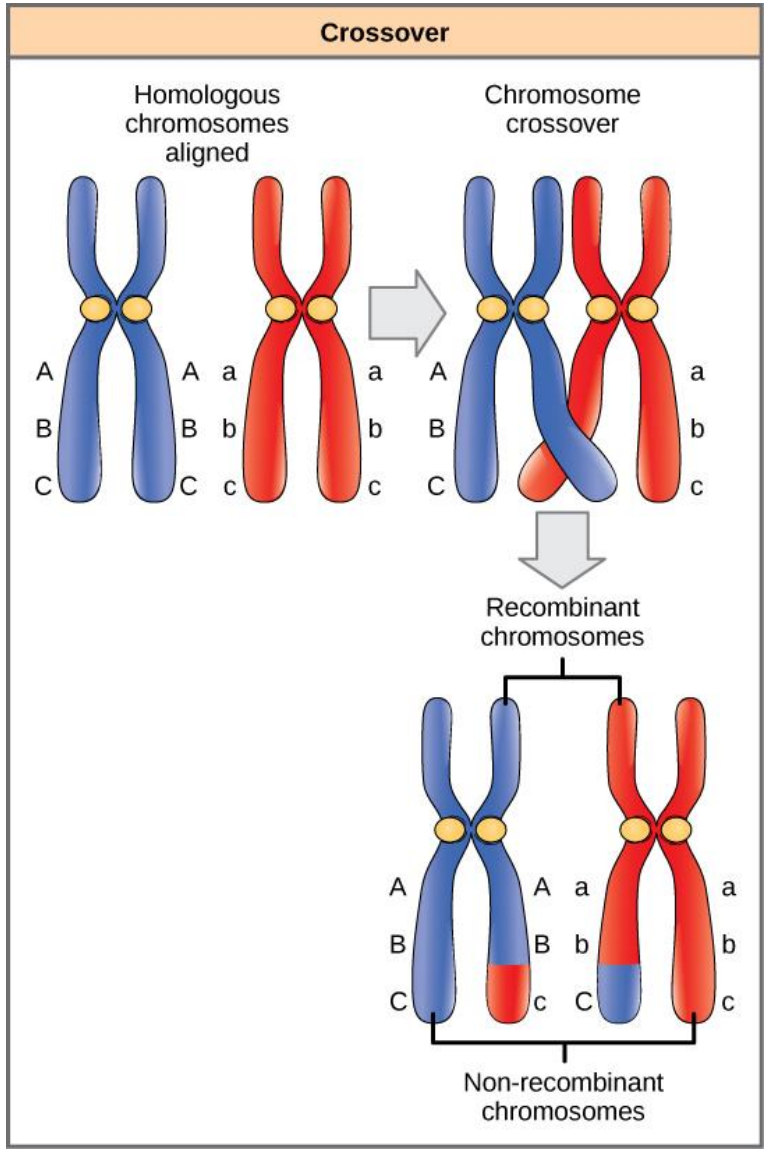


Nits and grits of recombination

- Similar mechanism with double-strand break repair (DSBR)
- double-Holliday junction (DHJ) intermediate, released through enzymatic cut (resolvase) producing either crossing over (CO) or non-crossing over (NCO) depending on which strand the cut occurs

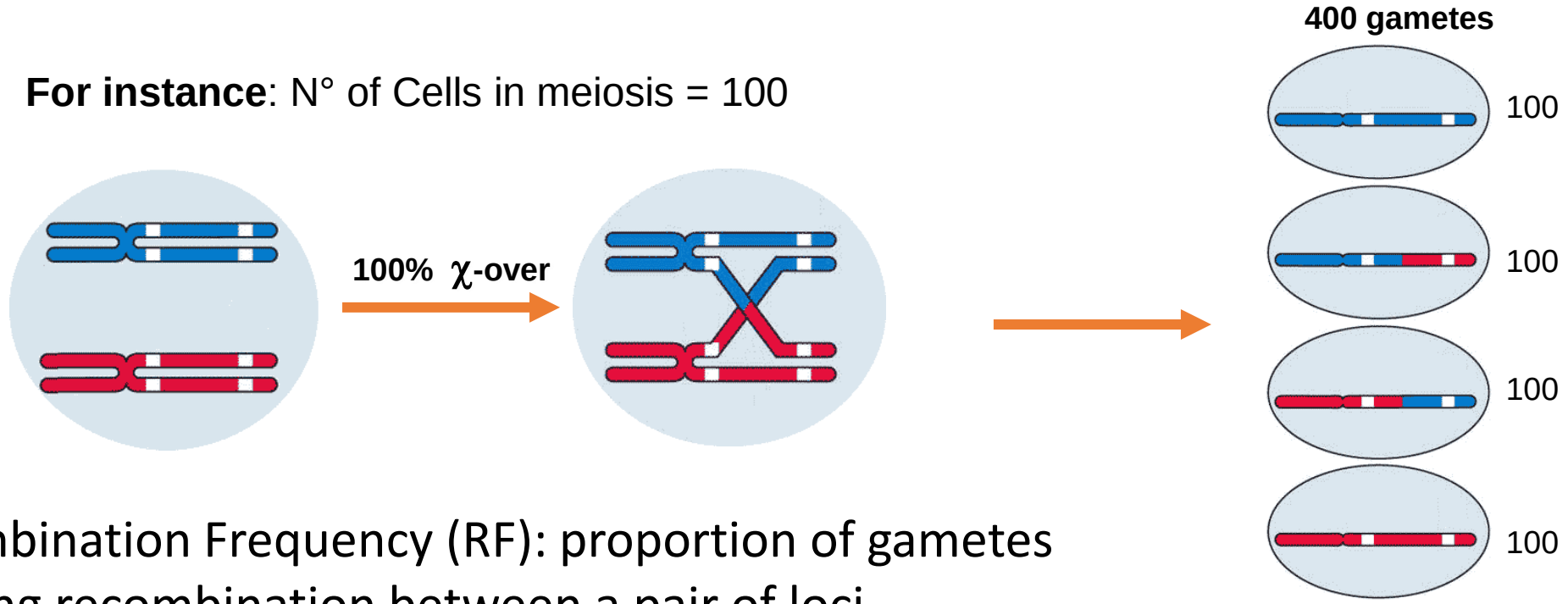
https://www.youtube.com/watch?v=3qgBKrAZCLg

The maximum value of recombination between any two loci is?



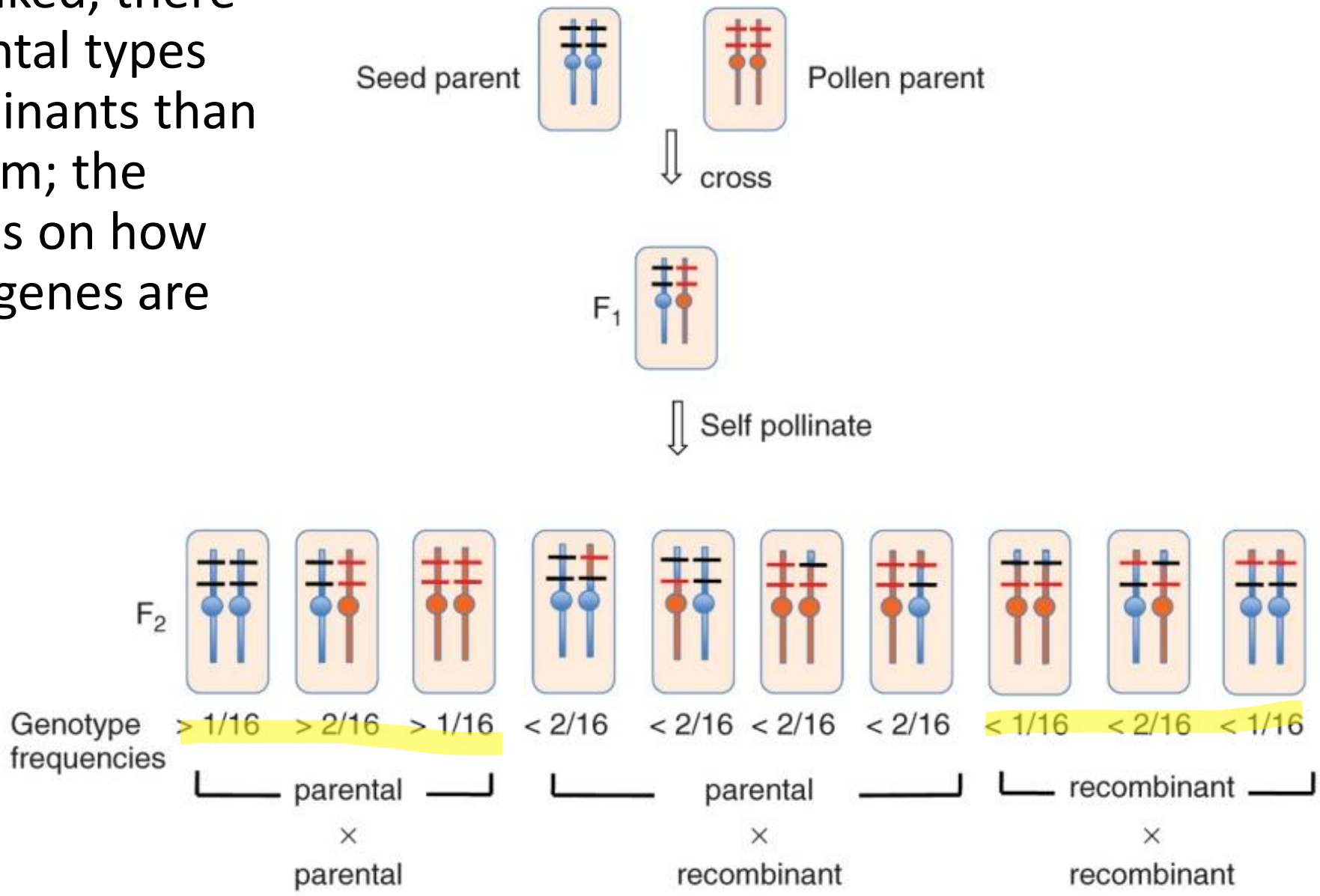
The maximum value of recombination between any two loci is?

Easy → random assortment of loci generates 50% recombination (non-linked loci produce 1:1 parental to non-parental ratio)



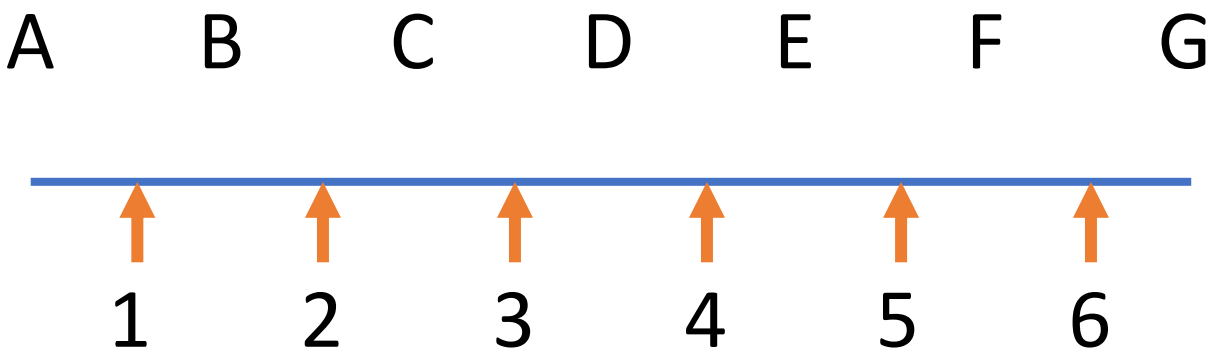
- Recombination Frequency (RF): proportion of gametes showing recombination between a pair of loci
- If genes are linked, you will see a higher percentage of parental gametes, making the RF < 0.50

If the genes are linked, there will be more parental types and fewer recombinants than expected at random; the frequency depends on how closely linked the genes are



RF is a function of physical distance

- At a first approximation, crossing over events take place at random positions along the chromosomes
- Consequently, the further two loci are apart, the more likely that there will be a crossing over event between them.

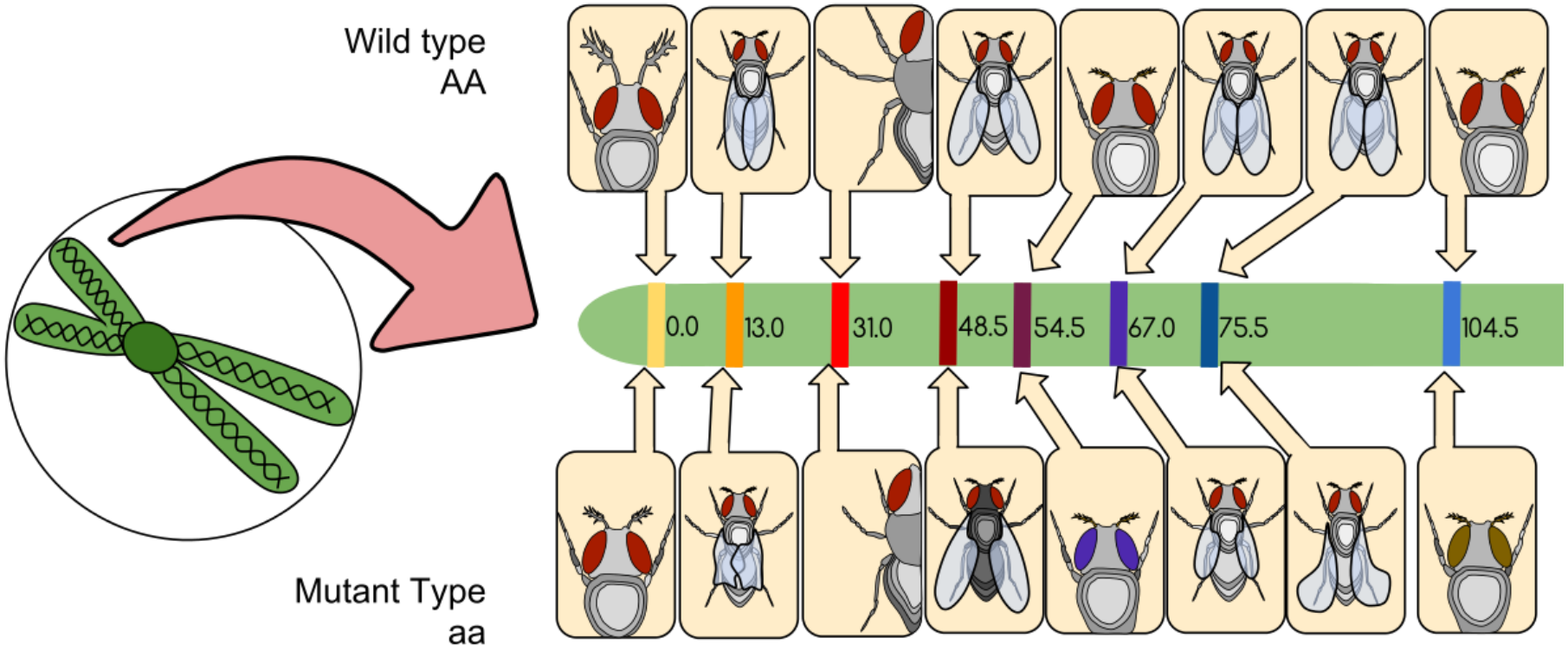


(actually, RF is not even across the chromosome length)

Imagine this as a rope and crossing over as a knot; every time you make a knot, A and G will recombine; but, e.g. D and E will recombine only if you make a knot in position 4

Linkage maps use RF to estimate positions of loci

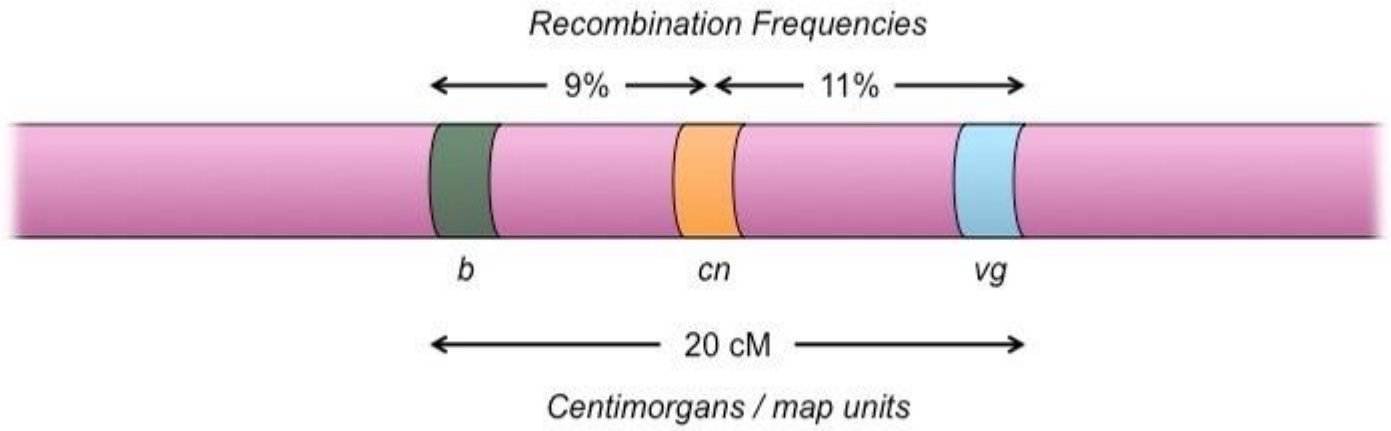
- Early genetic works, before DNA sequencing was a thing (but even after that) cleverly used RF between loci to create genetic maps of chromosomes
- Maps define the linear relationship between loci, and do so assessing RF between any pairs of loci



Alfred H. Sturtevant, Thomas Morgan, 1912

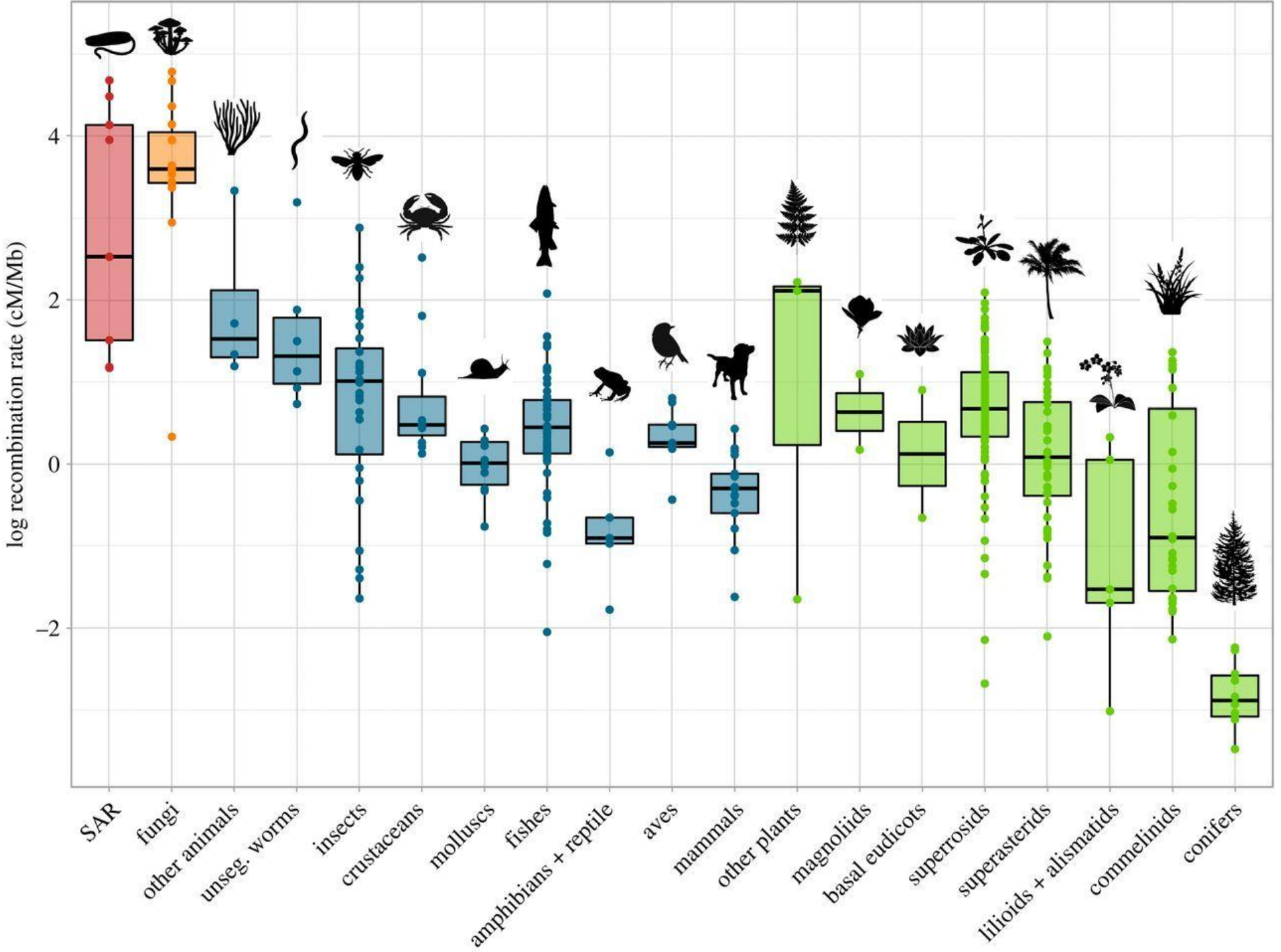
Unsurprisingly, genetic maps are measured in centiMorgans (cM)

- 1 cM equates to 1 observed recombination every 100 gametes
- 1 cM is equal to a 1% chance that two loci on a chromosome will become separated from one another due to a recombination event during meiosis



- Rule of thumb: 1cM corresponds to about 1Mb of DNA

The corresondance between physical map distance and genetic distance depends on the organism (as it depends on RF)

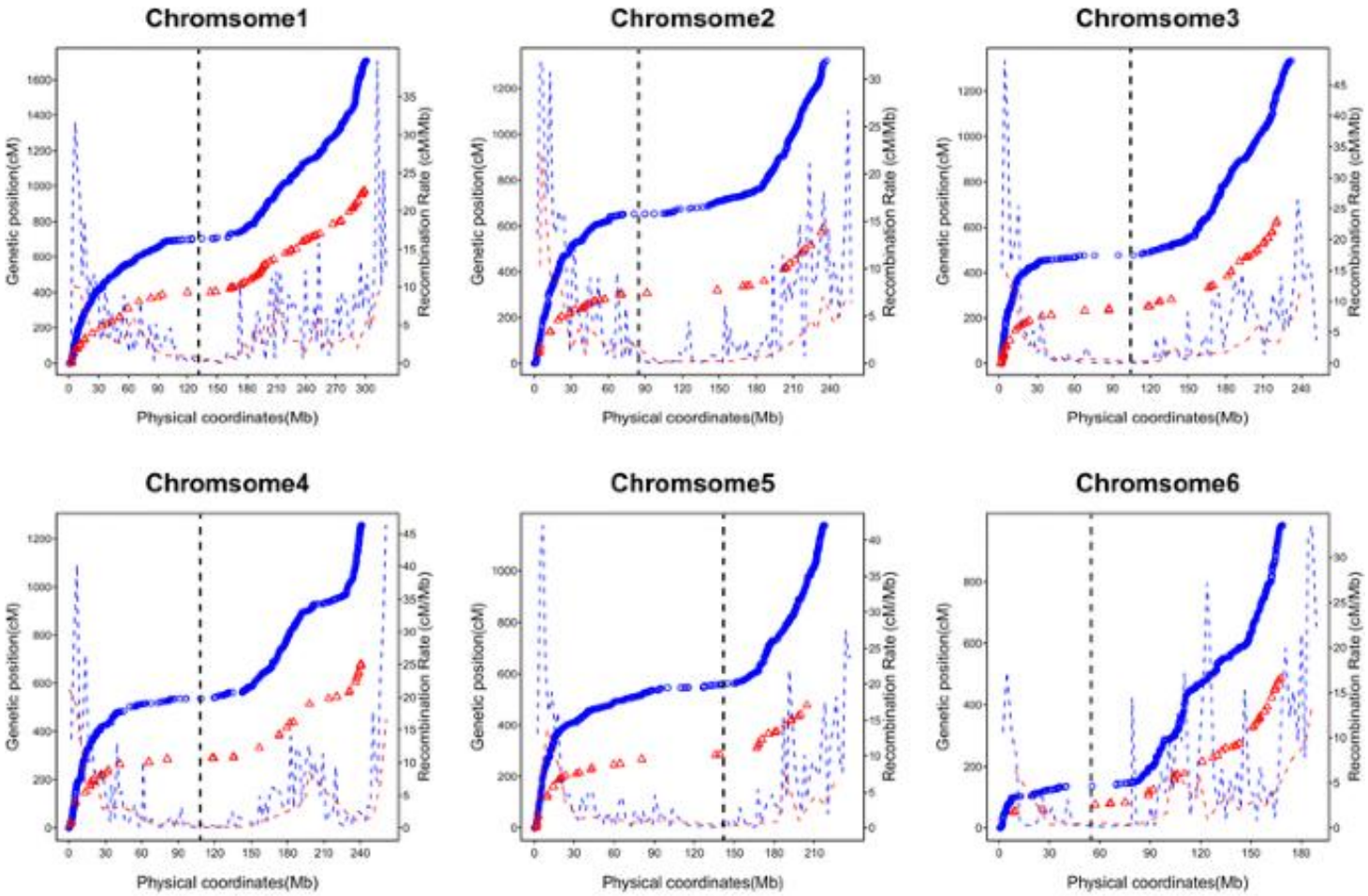


An ultra-high-density map as a community resource for discerning the genetic basis of quantitative traits in maize

Hongjun Liu, Yongchao Niu, Pedro J. Gonzalez-Portilla, Huangkai Zhou, Liya Wang, Tao Zuo, Cheng Qin, Shuaishuai Tai, Constantin Jansen, Yaou Shen, Haijian Lin, Michael Lee, Doreen Ware, Zhiming Zhang ✉, Thomas Lübberstedt ✉ & Guangtang Pan ✉

BMC Genomics **16**, Article number: 1078 (2015) | Cite this article

Map distance depends also on position in the chromosome



Factors influencing recombination rate

Chromosomal Structure & Features

- Centromeres & Telomeres: Recombination is generally suppressed near centromeres and increased near telomeres. Crossovers too close to the centromere can **disrupt kinetochore function**, which can lead to mis-segregation (e.g. aneuploidy)
- Heterochromatin vs Euchromatin: Recombination is lower in heterochromatin (condensed, gene-poor regions) and higher in euchromatin (gene-rich, open regions)

Sequence Motifs & Hotspots

- Recombination occurs more often at recombination hotspots, which are short DNA sequences where recombination initiates more frequently. The machinery that initiates recombination (e.g. SPO11 in many organisms) forms fewer DSBs near centromeres due to specific motifs (or lack thereof)
- GC-rich regions tend to have higher recombination rates
- Repetitive elements and transposons can either promote or suppress recombination depending on their type and context

Epigenetic Marks

- Histone modifications and DNA methylation can influence recombination by making regions more or less accessible to the recombination machinery

Developmental and Sex Differences

- In many species, recombination rates differ between males and females (heterochiasmy)
- The rate can also vary between meiosis I vs II, or depending on environmental stress during development

Evolutionary and Population Factors

- Regions under strong selection may suppress recombination (e.g., supergenes, centromeres)
- Recombination can be seen more frequently in regions where genetic diversity is advantageous

Genetic maps are important for a number of reasons:

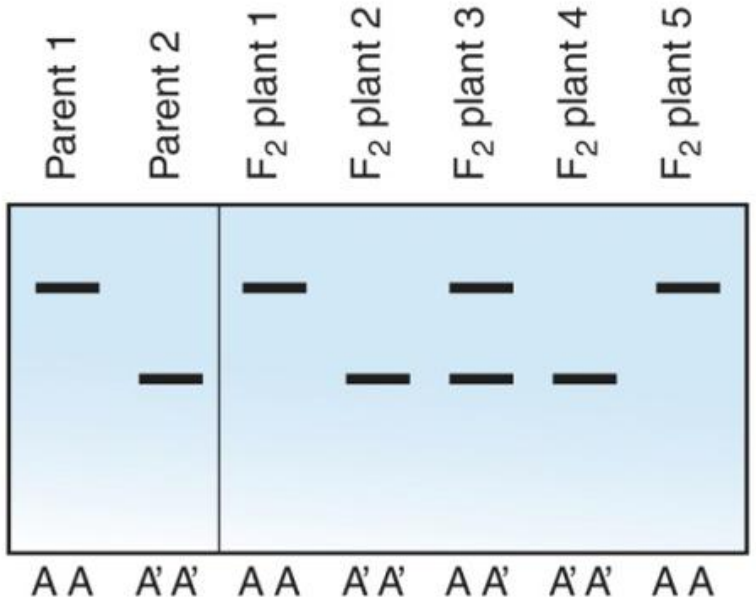
- Understand genome organization and topology
- Link function to position on the genome (through markers) via forward and reverse genetics
- <u>Anchor and orient sequencing data</u>

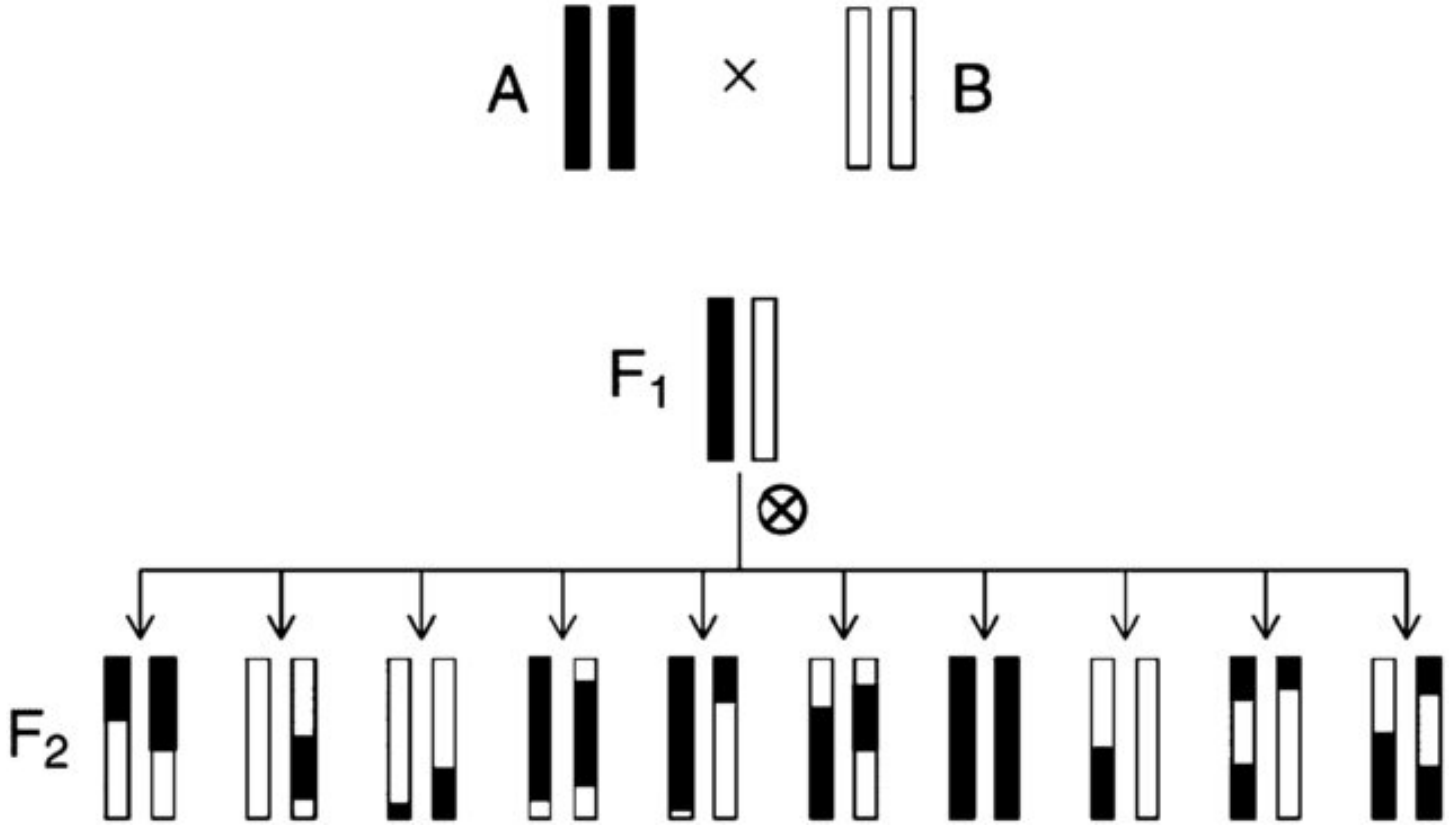
Genetic maps are the results of experiments

1. Identification of polymorphisms (molecular markers generation)
2. Selection of parental genotypes and breeding
3. Production of a segregating population (F2 – Backcross – etc.)
4. Genotyping of single individuals in the population to detect alleles at polymorphisms
5. Analyze segregation data and understand relation between loci
6. Combine linkage groups to reconstruct chromosomes

F2 populations

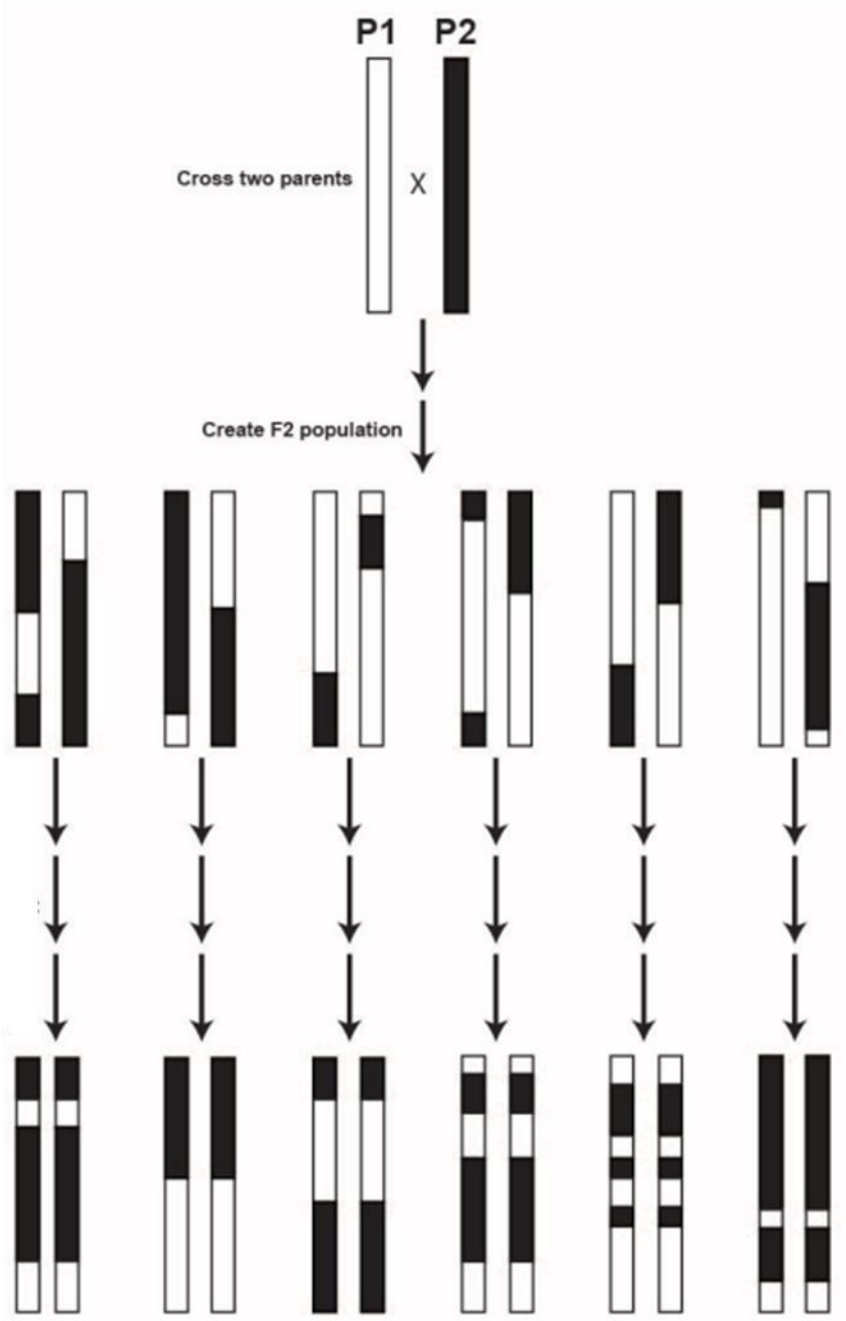
- All heterozygous loci segregate in a single meiosis
- Quick and easy





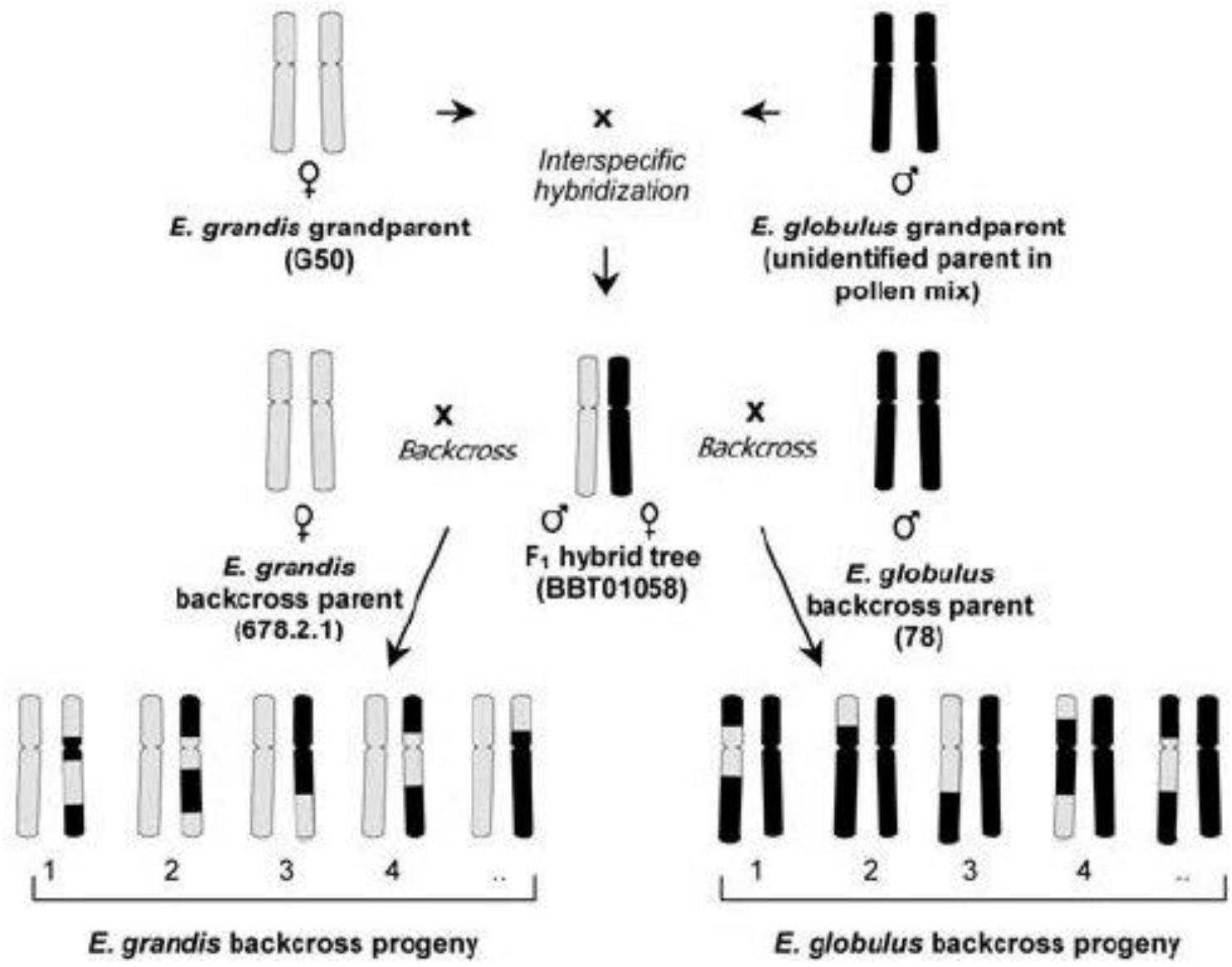
Recombinant inbred lines (RILs)

- Following the F2, subsequent selfing generations halve heterozygosity at each generation until fixation
- Intermating generation can be added to increase recombinations
- Recombination events are immortalized



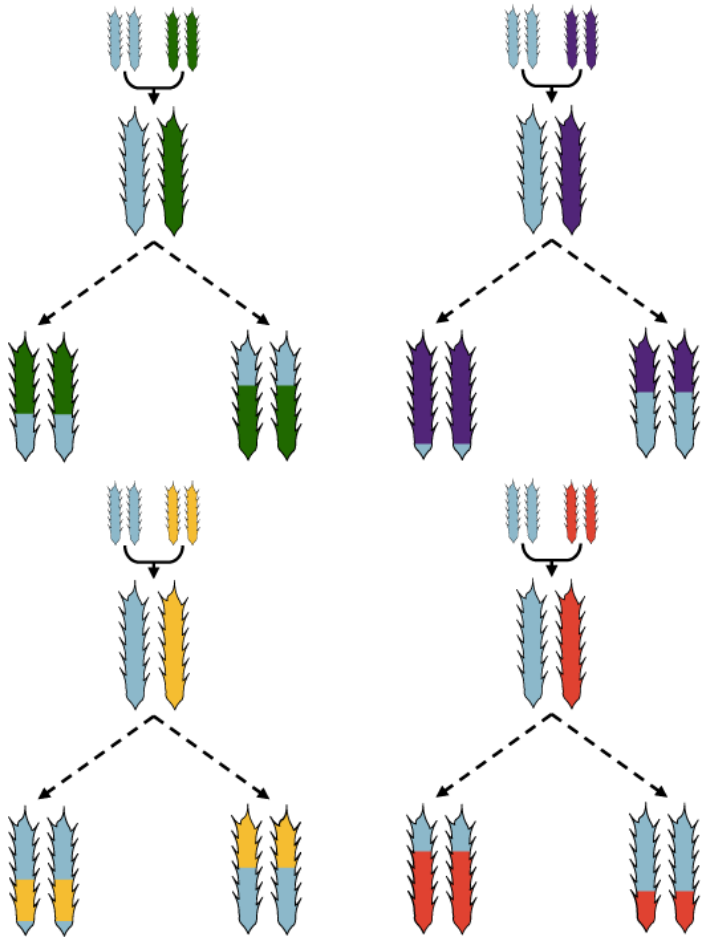
Backcross

- Progenies are crossed back with one of the parents
- Only the allele of the NON recurrent parent segregate
- Useful to study e.g. the introgression of a particular locus in a genome of interest

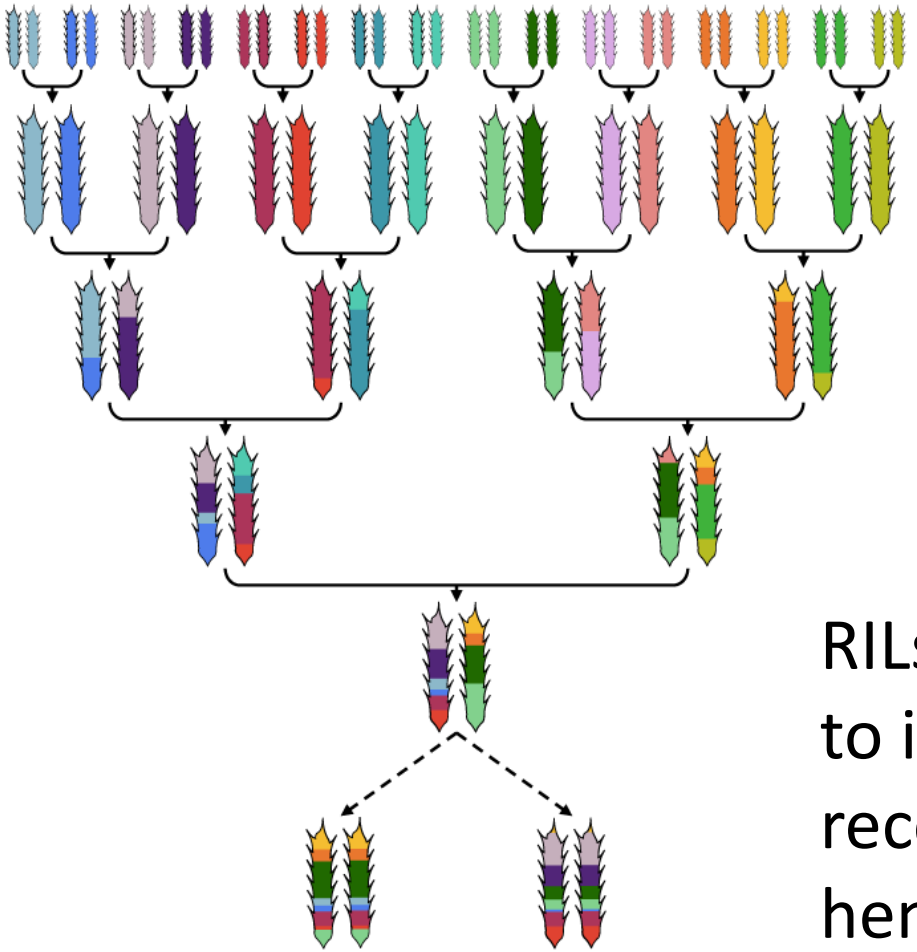


Multiparent populations

Nested Association Mapping (NAM) panel



Multi-parent Advanced Generation Inter-Cross (MAGIC)



RILs with extra steps to increase recombinations and hence map definition

How to produce a map

- Assume you have a trihybrid (three genes heteorizygous) that produces the following gametes:

Phenotype (or genotype)	# observed
ABC	46
Abc	63
ABc	4
aBc	381
abc	38
aBC	71
abC	2
AbC	395
TOT	**1000**

Step 1: reorder the gametes by frequency

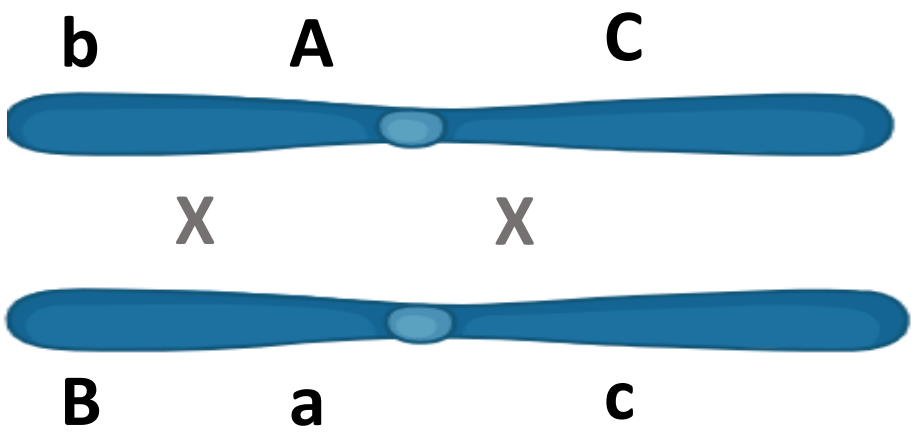
Step 2: look for parentals (= no recombinats) and look for double recombinants

Phenotype (or genotype)	# observed
AbC	395
aBc	381
aBC	71
Abc	63
ABC	46
abc	38
ABc	4
abC	2
TOT	**1000**

Step 3: find the middle locus

- Parentals: AbC, aBc
- Double recombinants: ABc, abC

Look at loci in pairs: AB, ab are not in parentals; Bc, bC are! If these are double recombinants, then I am betting that **Aa** is in the middle



Phenotype (or genotype)	# observed
AbC	395
aBc	381
aBC	71
Abc	63
ABC	46
abc	38
ABc	4
abC	2
TOT	**1000**

Step 4: calculate map distances

Pick a pair of loci and sum all the recombinants

- Say **Aa**, **Cc**: 71+63+4+2 = **140**

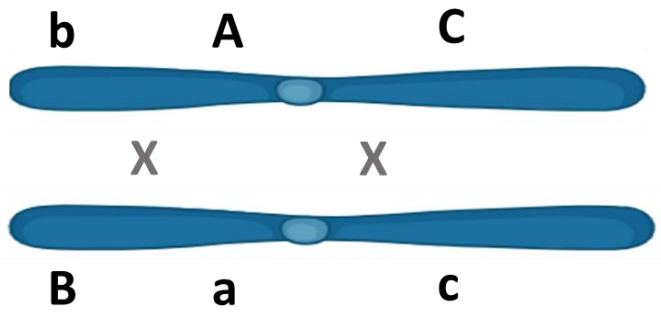
Get frequency over total number of gametes, and multiply by 100 to get cM

- 140/1000 = 0.140*100 = **14 cM**

Repeat for the other intervals:

- For **Bb**, **Aa:** (46+38+4+2)/1000*100= **9 cM**
- The distance between **Bb** and **Cc** must be **23 cM**

Phenotype (or genotype)	# observed
AbC	395
aBc	381
aBC	71
Abc	63
ABC	46
abc	38
ABc	4
abC	2
TOT	**1000**



Plant Biotechnology Journal

Plant Biotechnology Journal (2015) **13**, pp. 648–663 doi: 10.1111/pbi.12288

A high-density, SNP-based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding

Marco Maccaferri[1,*], Andrea Ricci[1], Silvio Salvi[1], Sara Giulia Milner[1], Enrico Noli[1], Pier Luigi Martelli[2], Rita Casadio[2], Eduard Akhunov[3], Simone Scalabrin[4,5], Vera Vendramin[4,5], Karim Ammar[6], Antonio Blanco[7], Francesca Desiderio[8], Assaf Distelfeld[9], Jorge Dubcovsky[10,11], Tzion Fahima[12], Justin Faris[13], Abraham Korol[12], Andrea Massi[14], Anna Maria Mastrangelo[15], Michele Morgante[4,5], Curtis Pozniak[16], Amidou N'Diaye[16], Steven Xu[13] and Roberto Tuberosa[1]

Summary

Consensus linkage maps are important tools in crop genomics. We have assembled a high-density tetraploid wheat consensus map by integrating 13 data sets from independent biparental populations involving durum wheat cultivars (*Triticum turgidum* ssp. *durum*), cultivated emmer (*T. turgidum* ssp. *dicoccum*) and their ancestor (wild emmer, *T. turgidum* ssp. *dicoccoides*). The consensus map harboured 30 144 markers (including 26 626 SNPs and 791 SSRs) half of which were present in at least two component maps. The final map spanned 2631 cM of all 14 durum wheat chromosomes and, differently from the individual component maps, all markers fell within the 14 linkage groups. Marker density per genetic distance unit peaked at centromeric regions, likely due to a combination of low recombination rate in the centromeric regions and even gene distribution along the chromosomes. Comparisons with bread wheat indicated fewer regions with recombination suppression, making this consensus map valuable for mapping in the A and B genomes of both durum and bread wheat. Sequence similarity analysis allowed us to relate mapped gene-derived SNPs to chromosome-specific transcripts. Dense patterns of homeologous relationships have been established between the A- and B-genome maps and between nonsyntenic homeologous chromosome regions as well, the latter tracing to ancient translocation events. The gene-based homeologous relationships are valuable to infer the map location of homeologs of target loci/QTLs. Because most SNP and SSR markers were previously mapped in bread wheat, this consensus map will facilitate a more effective integration and exploitation of genes and QTL for wheat breeding purposes.

- Genetic maps depend on experiments (e.g. genotypes, markers) used to derive them
- Maps can be combined across experiments

Table 1 Details of the 13 mapping populations and of the corresponding single component maps

Mapping populations					Molecular markers					Linkage group		
					Genomic		Illumina			Linkage	Total	Intermarker
		Contributing	Type	Size	SSR	DArT®	SNP†	Others‡	Total	group	length	distance§
Parents	Acronym	Institution*	DH/RIL	no.	no.	no.	no.	no.	no.	no.	cM	cM/marker
T. durum × *T. durum*												
Colosseo × Lloyd¶	Cl × Ld	UNIBO/UNIUD/PSB	RIL	176	184	372	6163	1227	7946	20	2063.9	0.33
Meridiano × Claudio¶	Mr × Cd	UNIBO/UNIUD/PSB	RIL	181	178	608	5097	87	5970	27	2238.8	0.43
Simeto × Levante¶	Sm × Lv	UNIBO/UNIUD/PSB	RIL	180	142	335	5315	6	5798	30	2184.7	0.40
Mohawk × Cocorit69	Mh × Cr	CIMMYT/USASK	RIL	81	–	–	5554	–	5554	31	2012.7	0.36
Svevo × Ciccio¶	Sv × Cc	UNIBA	RIL	103	16	213	5246	12	5487	26	1887.6	0.36
W9292-260D3 × Kofa	G9586	AAFC/USASK	DH	155	34	–	3676	2	3712	33	1685.0	0.46
Kofa × Svevo¶	Kf × Sv	UNIBO/UNIUD/PSB	RIL	249	205	–	–	38	243	18	1256.2	–
Kofa × UC1113¶	Kf × UC	UCDAVIS	RIL	93	172	–	–	31	203	24	755.1	–
T. durum × *T. dicoccum*												
Ben × PI41025	Bn × PI_41025	USDA-ARS	RIL	200	111	–	2456	–	2567	14	2526.9	–
Simeto × Molise Colli	Sm × Ml	CRA-Foggia	RIL	136	26	–	8926	–	8952	15	3028.4	0.34
Latino × MG5323	Lt × MG_5323	CRA-Fiorenzuola/UNIBA	RIL	82	216	–	10 572	23	10 811	14	2363.4	0.23
T. durum × *T. dicoccoides*												
Langdon × G18-16¶	Ln × G18-16	UHAIFA	RIL	152	120	148	–	–	268	20	1577.3	–
Svevo × Zavitan	Sv × Zv	UTELAVIV	RIL	140	–	–	10 911	–	10 911	14	2258.0	0.20

RIL, recombinant inbred line; DH, double haploid; SSR, simple sequence repeat; DArT®, Diversity Array Technology; SNP, single nucleotide polymorphism.

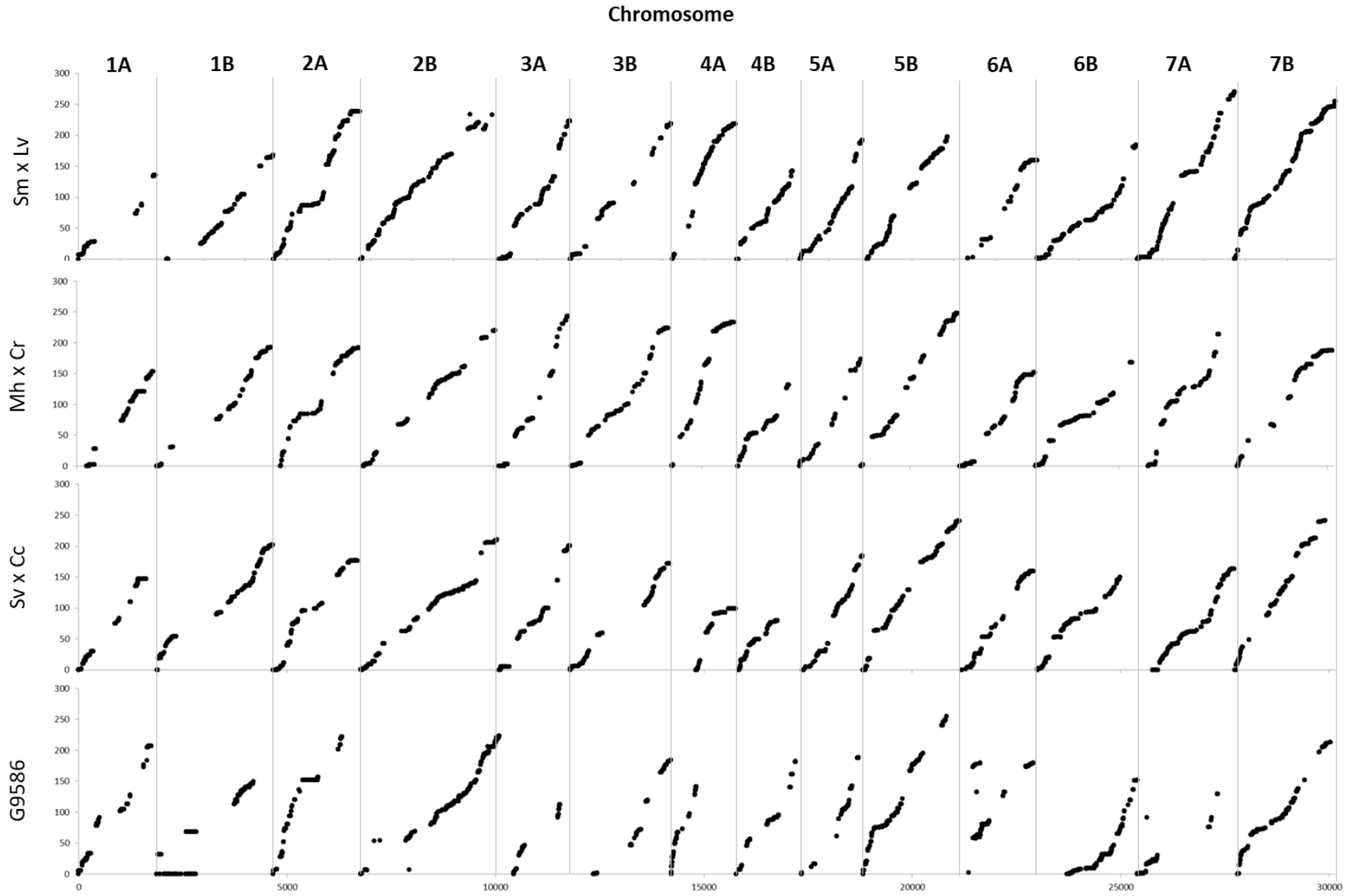
*UNIBO, University of Bologna; UNIUD, University of Udine; PSB, Produttori Sementi Bologna; CIMMYT, International Maize and Wheat Improvement Center; USASK, University of Saskatchewan; UNIBA, University of Bari; AAFC, Agriculture and Agri-Food Canada; USDA-ARS, Cereal Crop Research Unit, Fargo; CRA, Consiglio per la Ricerca e la Sperimentazione in Agricoltura; UHAIFA, University of Haifa; UTELAVIV, Tel Aviv University.

†Illumina iSelect 90K wheat SNP array used for 9 populations; Illumina iSelect 9K wheat SNP array used for Bn × PI41025.

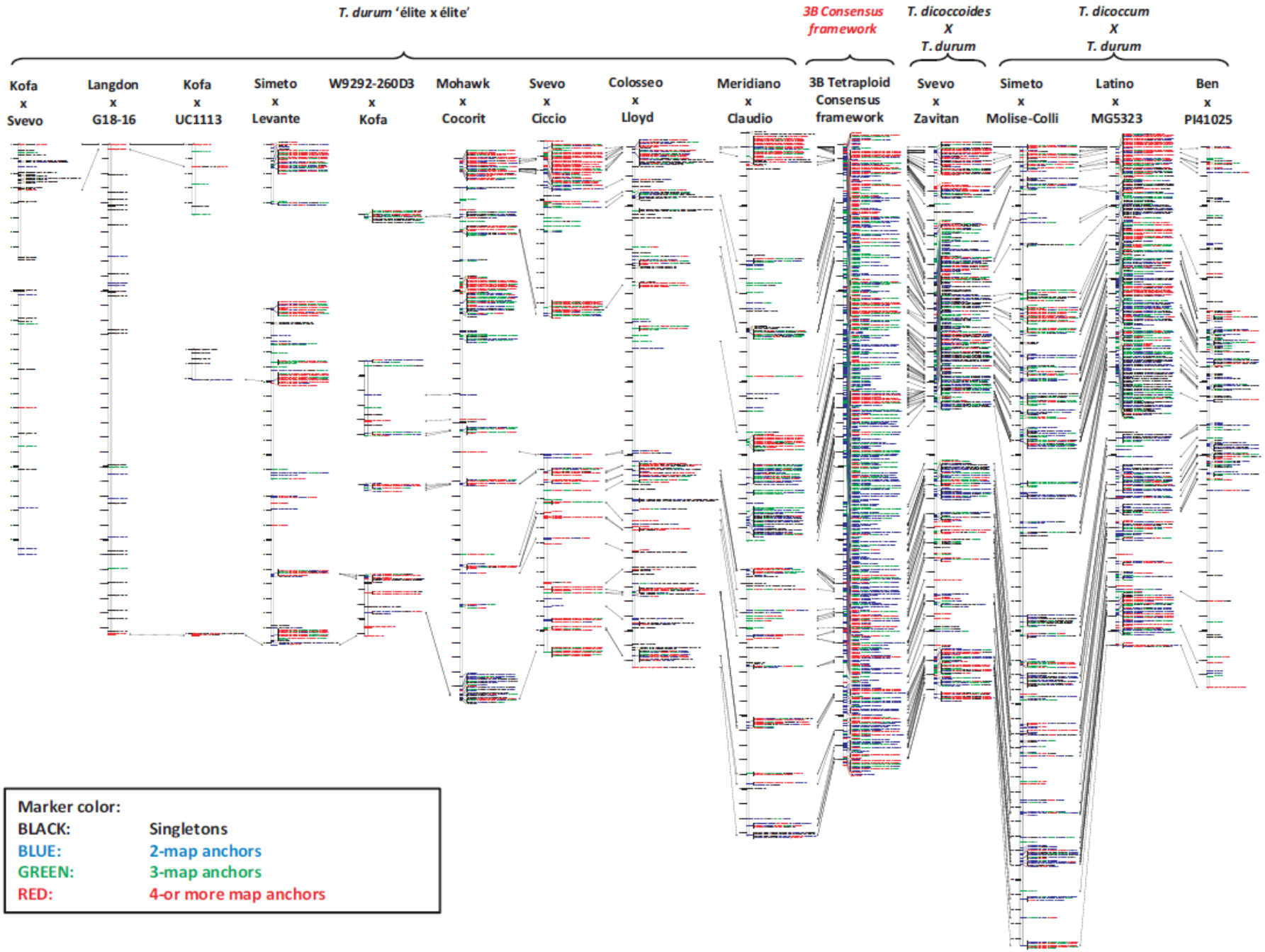
‡Include sequence tagged sites, morphological and biochemical markers, and 1065 sequence-based genotyping SNPs for Cl × Ld.

Representation of chromosomes and also map lengths are dependent on markers and genetic materials used

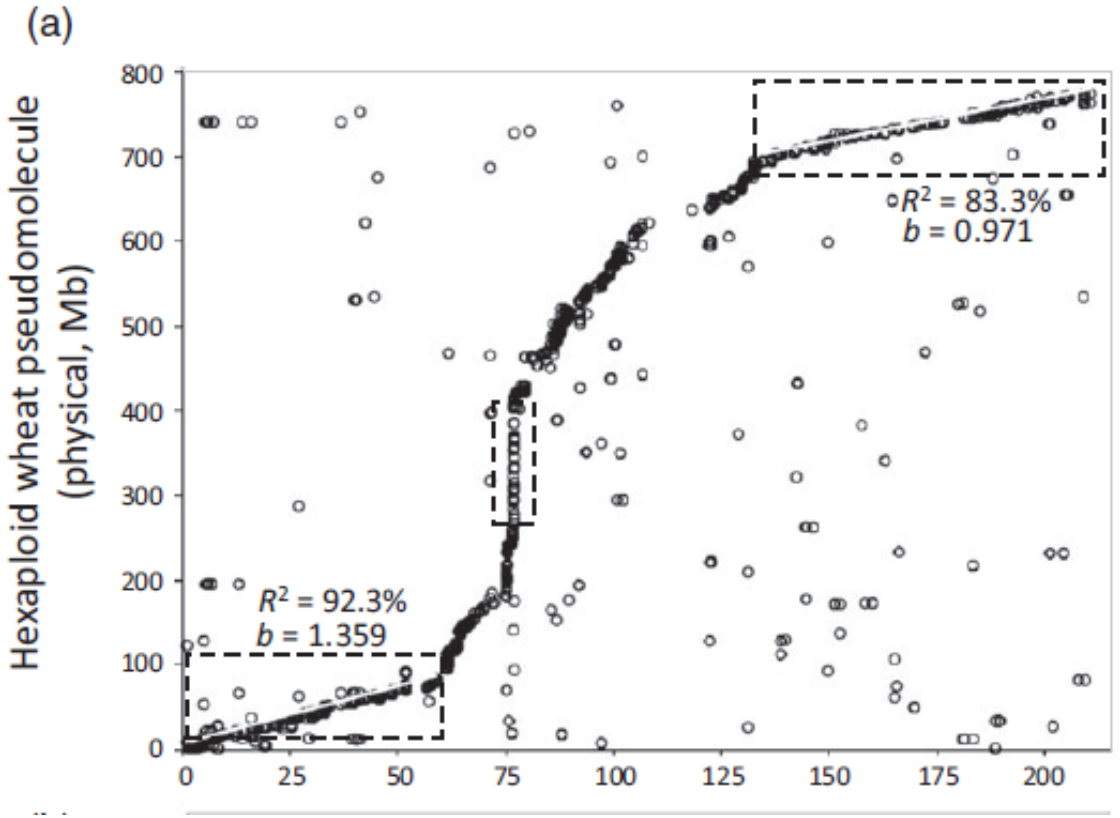
- <u>Genetic maps depend on experiments</u>



Maps can be combined across experiments to develop a «consensus map» considering overlap of markers across individual maps



The relation between physical distance (x axis) and genetic distance (y axis) is highly context dependent



Genetic maps are cool but…

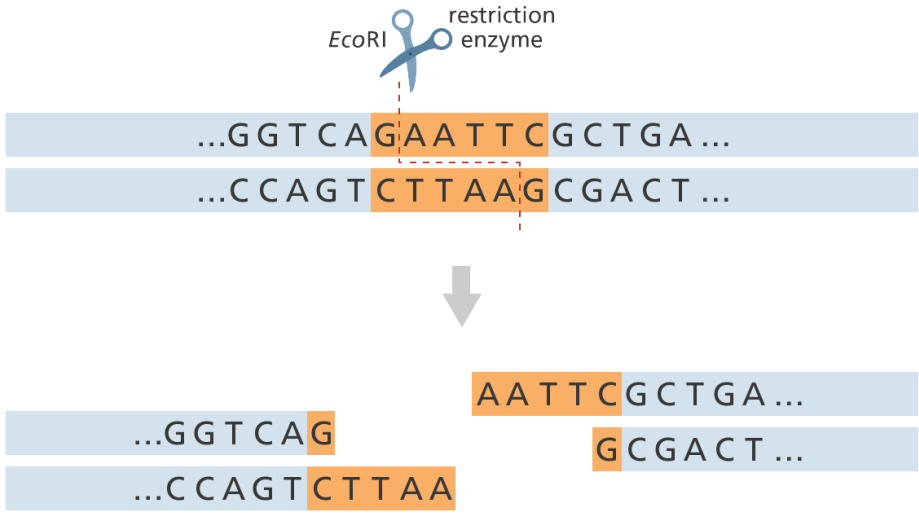
- The resolution of a genetic map depends on the number of crossovers that have been scored. Our rule of thumb is 1cM (1/100 gametes) corresponds to 1 Mb. To resolve the map to 0.01 Mb (ballpark gene size) would need observing 1 recombinant over 100,000 gametes
- Crossovers are not equally frequent throughout the genome, and this may lead to incorrect positioning of loci

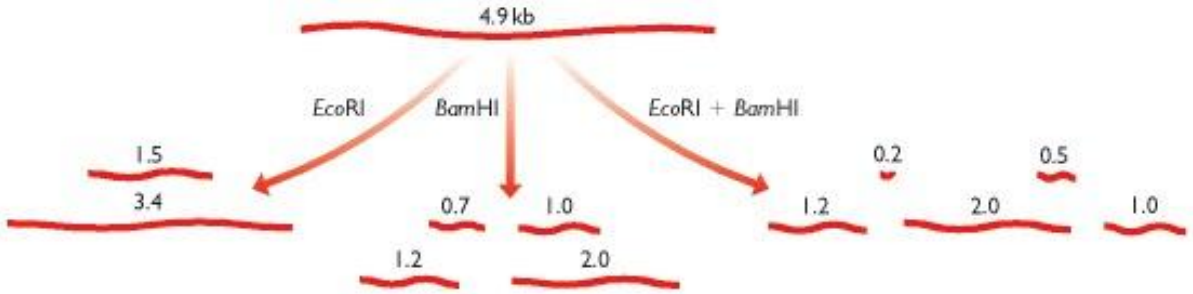
From genetic maps to physical maps

- Physical mapping gives an estimation of the (physical) distance between specific known DNA sequences on a chromosome
- The distance is expressed as the number of base pairs between them.
- **Restriction mapping** (AKA fingerprinting) has been a breakthrough technology to move from genetic maps to physical maps at the dawn of the genomic era

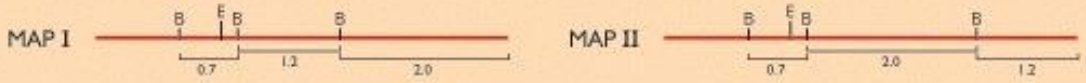
The idea is to map the location of restriction sites across the chromosomes (remember RFLP?)

- A physical map can be generated by aligning the different restriction maps along the chromosomes.





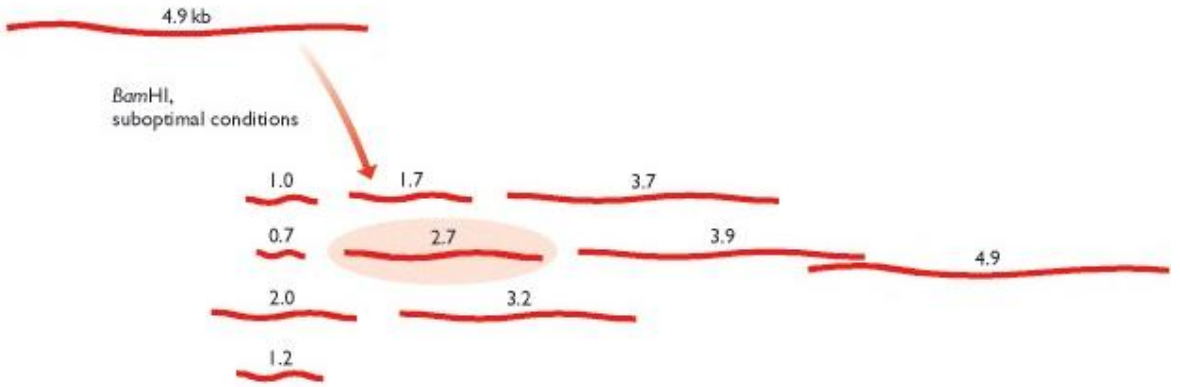
INTERPRETATION OF THE DOUBLE RESTRICTION

Fragments	Conclusions
0.2 kb, 0.5 kb	These must derive from the 0.7 kb *BamHI* fragment, which therefore has an internal *EcoRI* site: 
1.0 kb	This must be a *BamHI* fragment with no internal *EcoRI* site. We can account for the 1.5 kb *EcoRI* fragment if we place the 1.0 kb fragment thus: 
1.2 kb, 2.0 kb	These must also be *BamHI* fragments with no internal *EcoRI* sites. They must lie within the 3.4 kb *EcoRI* fragment. There are two possibilities: 

PREDICTED RESULTS OF A PARTIAL *BamHI* RESTRICTION

If Map I is correct, then the partial restriction products will include a fragment of 1.2 + 0.7 = 1.9 kb
If Map II is correct, then the partial restriction products will include a fragment of 2.0 + 0.7 = 2.7 kb

- Relative sizing of fragments is determined by combination of Res
- The exact distance and sizing can be determined by gel electrophoresis, using a standard
- Depending on the frequency of the RE size, you can tweak sizing and relation of fragments

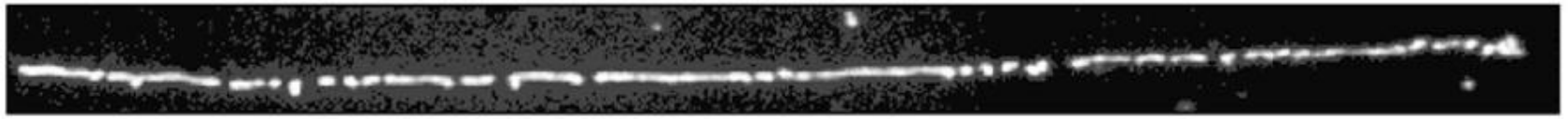


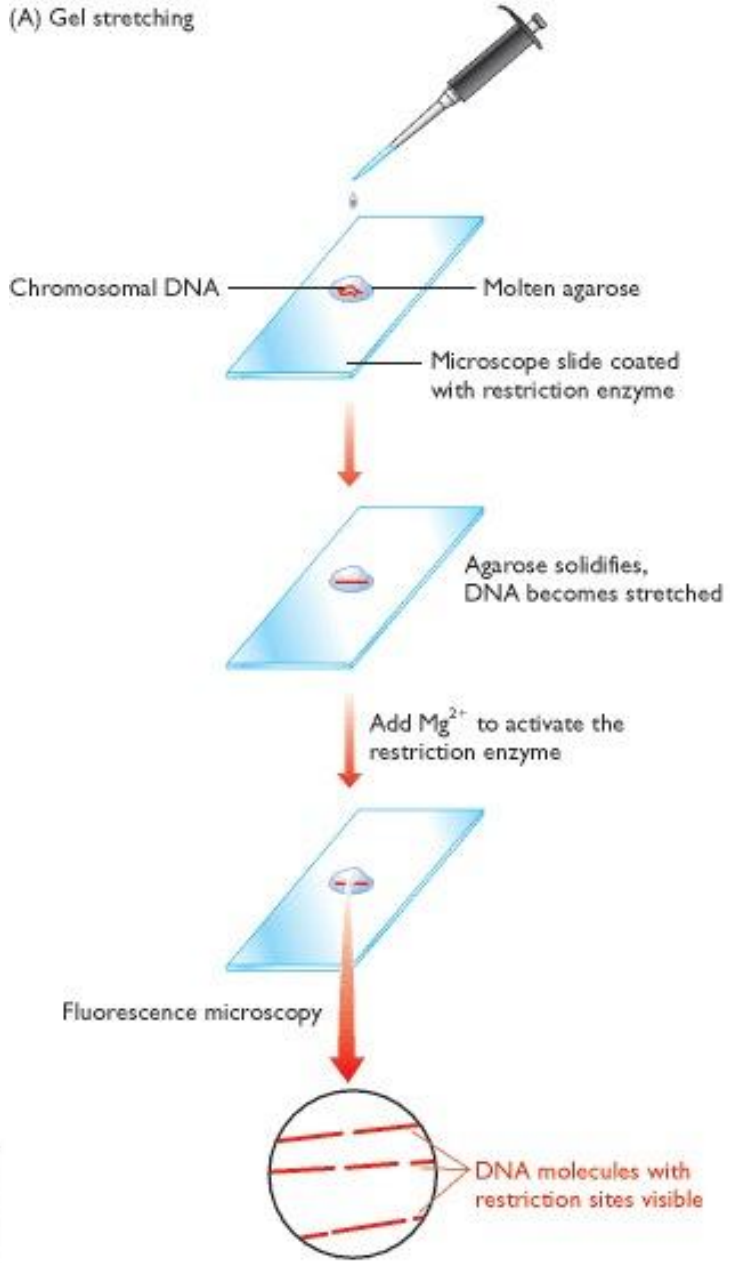
CONCLUSION
Map II is correct

Clearly, the whole thing becomes too complex when dealing with longer fragments

Optical mapping is another method to observe physical distance between loci, and is based on observation of the cut molecules with a microscope

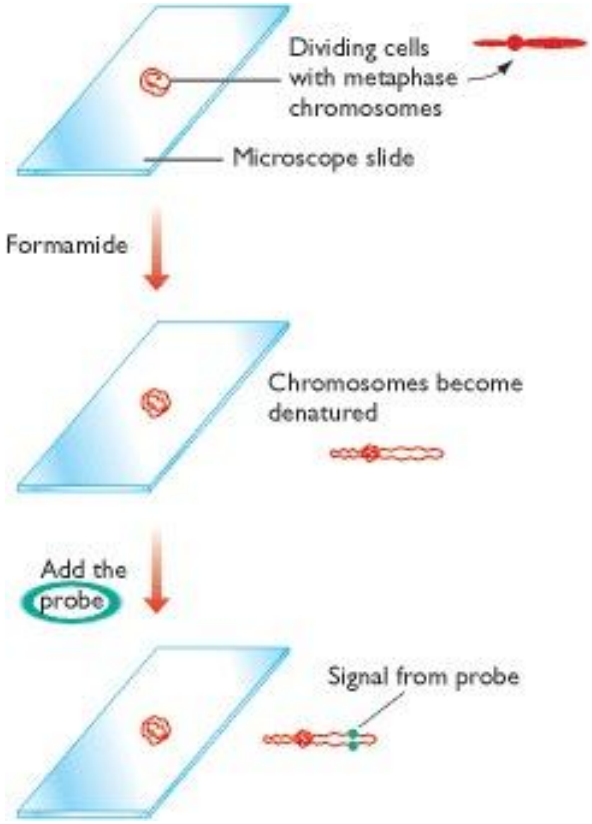
- You need to linearize molecules and attach them to a surface; one way to do so is to use gel stretching, a matrix which extends the DNA so that gaps caused by RE can be seen

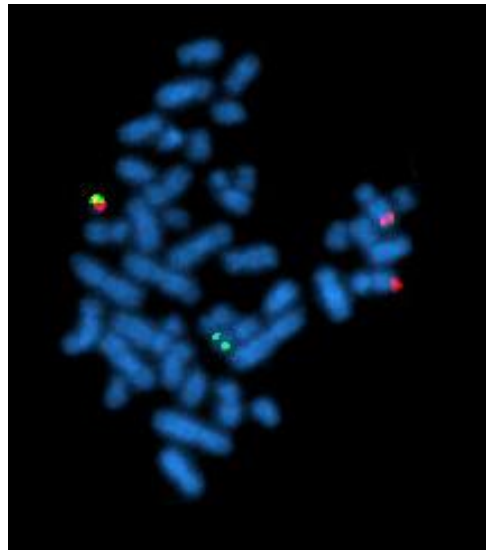




Fluorescent in-situ hybridization (FISH) is yet another method to achieve physical information on the localization of genomic loci

- In optical mapping, the marker is a restriction site and it is visualized as a gap in an extended DNA fiber. In FISH, the marker is a DNA sequence that is visualized by hybridization with a fluorescent probe





Good for chromosome-scale localization (including rearrangements)

Why are maps important for sequencing

- Genetic and physical maps represent the first description of a complex genome
- Maps are useful to order genetic markers; genetic markers are anchors to associate physical DNA fragments to chromosomes
- **Mapping is the first pillar for the production of a genome sequence; the second pillar being the capacity to read the nucleotide sequence of DNA fragments**

If you can *sequence* the DNA fragments that you ordered with maps, bingo! Genomes can be reconstructed (we will see how)