GENETICS AND MOLECULAR BIOLOGY FOR ENVIRONMENTAL ANALYSIS

MOLECULAR ECOLOGY

LESSON 8: MOLECULAR MARKERS IN ECOLOGY

Prof. Alberto Pallavicini
pallavic@units.it

MOLECULAR MARKERS

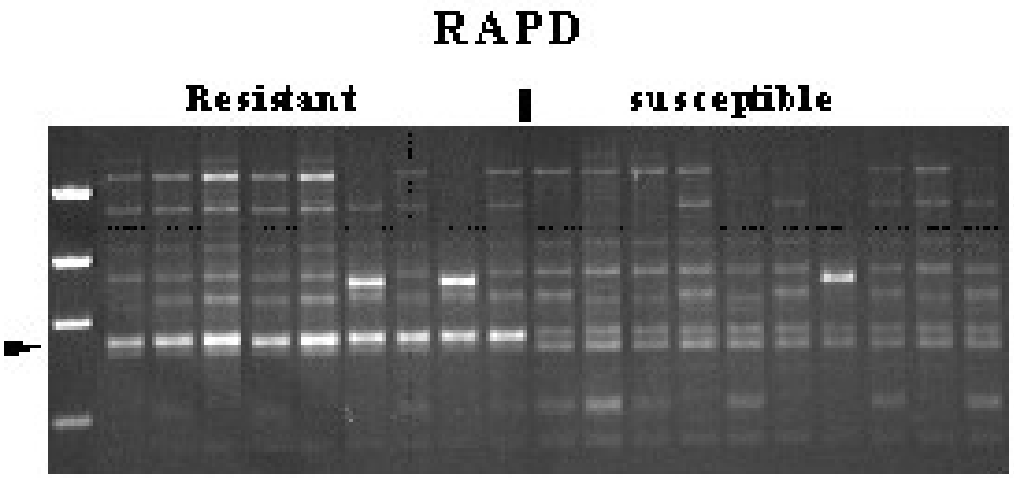
- An important point to remember is that molecular markers are simply tools that can be used for generating data; like anything else, the job can be done properly only if the correct tools are chosen.
- The other point to bear in mind while learning about molecular markers is that we are using them to answer ecological questions.

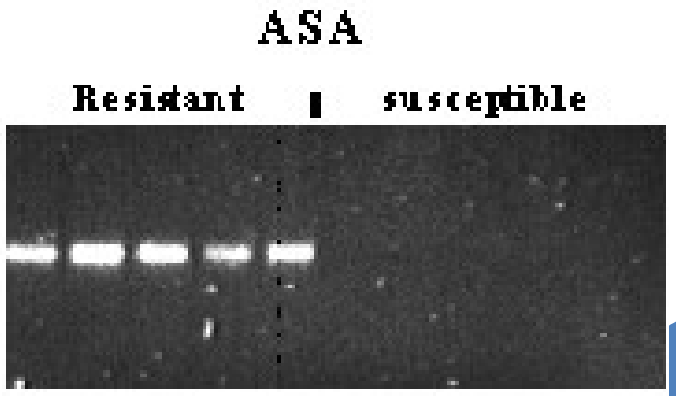
MOLECULAR MARKERS

- Several factors need to be taken into account when choosing a marker.
- The expected level of variability.
  - markers that allow us to differentiate between closely related organisms will need to be highly variable
  - Practical concerns

CO-DOMINANT AND DOMINANT MARKERS

- **Co-dominant** markers allow us to identify all of the alleles that are present at a particular locus, whereas **dominant** markers will reveal only a single dominant allele.

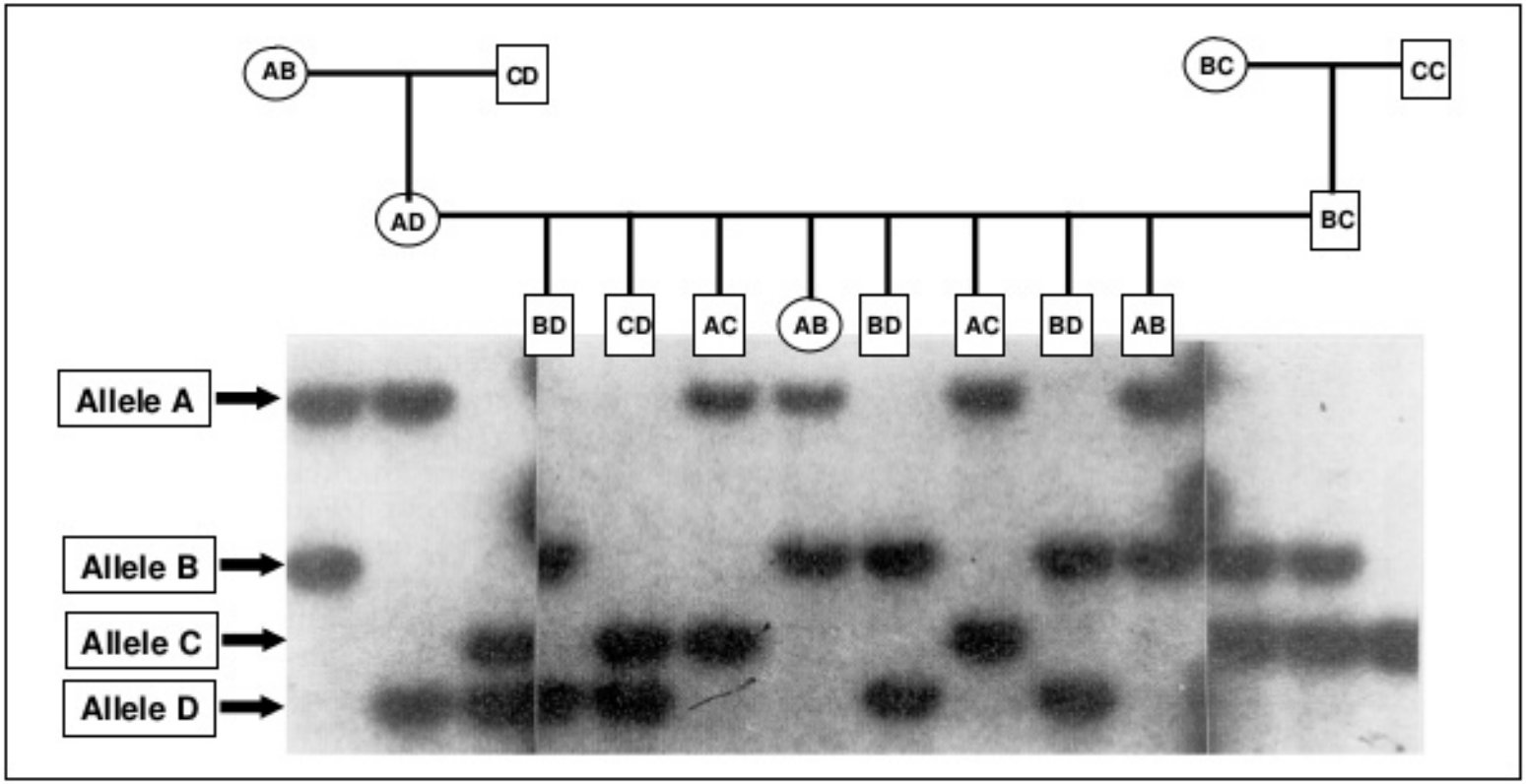




CO-DOMINANT MARKERS

- In a diploid species, each dominant marker will identify one allele in a homozygous individual and two alleles in a heterozygous individidual.
- This ability to distinguish between homozygotes and heterozygotes is one of the most important features of co-dominant markers because it means that we can calculate easily the allele frequencies for pooled samples (such as populations).

Microsatellites are co-dominant markers



- each co-dominant marker characterizes a single locus, most projects will use multiple co-dominant markers to generate data from a number of different loci so that conclusions are not based on a single, possibly atypical, locus.

TRADITIONAL VIEW OF MOLECULAR MARKERS

- This list of traditional molecular markers including their acronyms can give us some idea of the basis of the molecular marker. It also describes the differences (polymorphisms) in DNA sequences they target.
- **Isozyme**: By measuring variations in enzymes, isozyme analysis exploits differences in the genes that code for or regulate enzyme synthesis or activity.
-

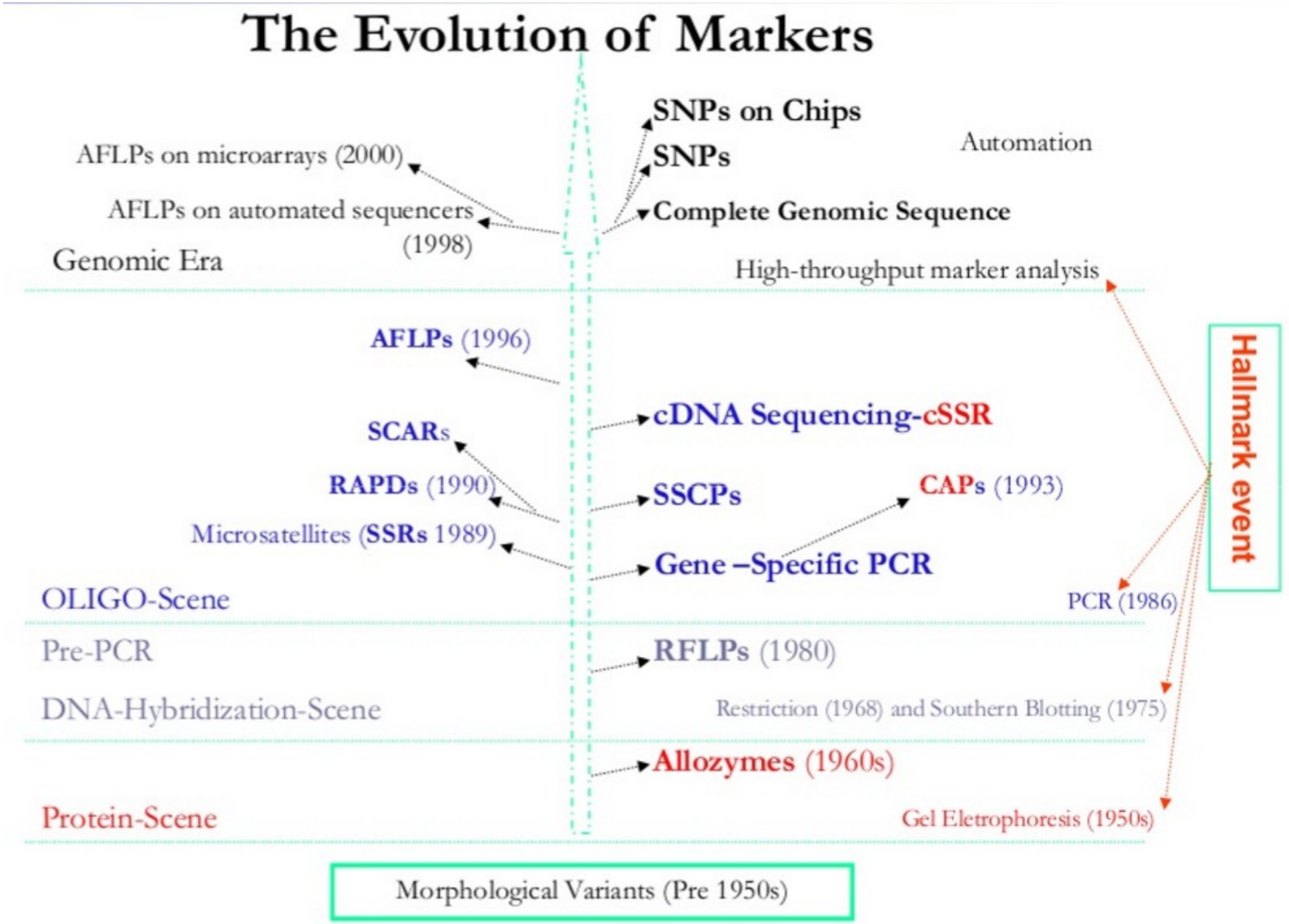
TRADITIONAL VIEW OF MOLECULAR MARKERS

RFLP (Restriction Fragment Length Polymorphism): Indirectly measure DNA sequence differences based upon the varying lengths of DNA fragments resulting from cutting it with restriction enzymes. These "fragment length polymorphisms" are visualized by hybridizing the cut DNA with labeled probes from DNA libraries.

RAPD (Random Amplified Polymorphic DNA): Utilizing a large number of short DNA primers with varying sequences, this technique exploits differences in the primer binding sites as different DNA will be amplified by the

TRADITIONAL VIEW OF MOLECULAR MARKERS

- **AFLP (Amplified Fragment Length Polymorphism)**: Utilizing restriction enzymes and a large number of short DNA primers with varying sequences, this technique exploits differences in the primer binding sites as different DNA will be amplified using PCR.
- **SSR (Simple Sequence Repeat)** or microsatellite: Using PCR, this technique exploits differences in short repetitive sequences (e.g., CAA vs. CAACAACAA) by using specifically designed DNA primers that bind on each side of repetitive DNA sequences.

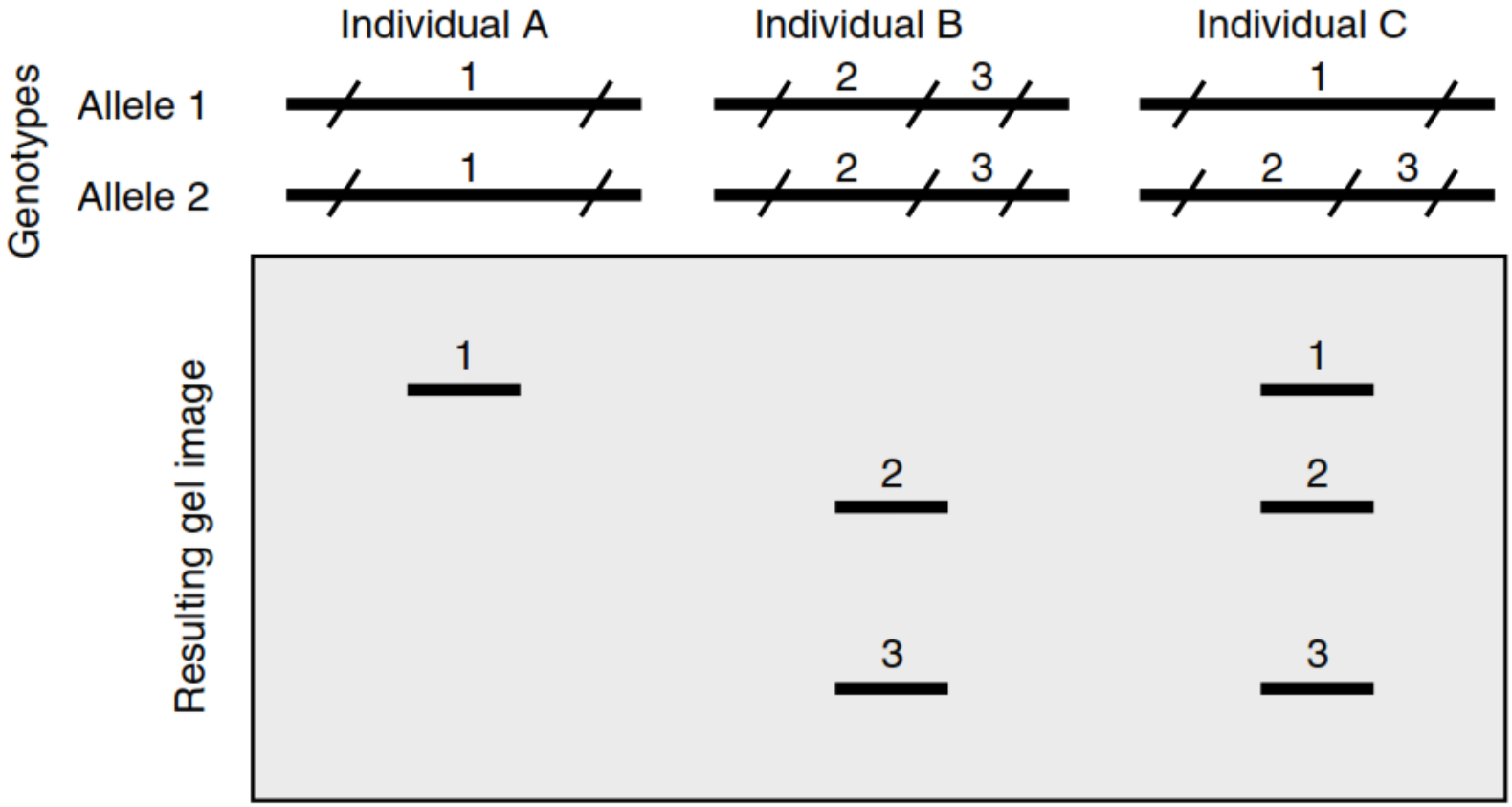

The Evolution of Markers
SNPs on Chips
SNPs
Automation
AFLPs on microarrays (2000)
AFLPs on automated sequencers (1998)
Complete Genomic Sequence
Genomic Era
High-throughput marker analysis
AFLPs (1996)
Hallmark event
cDNA Sequencing-cSSR
SCARs
RAPDs (1990)
SSCPs
CAPs (1993)
Microsatellites (SSRs 1989)
Gene –Specific PCR
OLIGO-Scene
PCR (1986)
Pre-PCR
RFLPs (1980)
DNA-Hybridization-Scene
Restriction (1968) and Southern Blotting (1975)
Allozymes (1960s)
Protein-Scene
Gel Eletrophoresis (1950s)
Morphological Variants (Pre 1950s)

ALLOZYMES

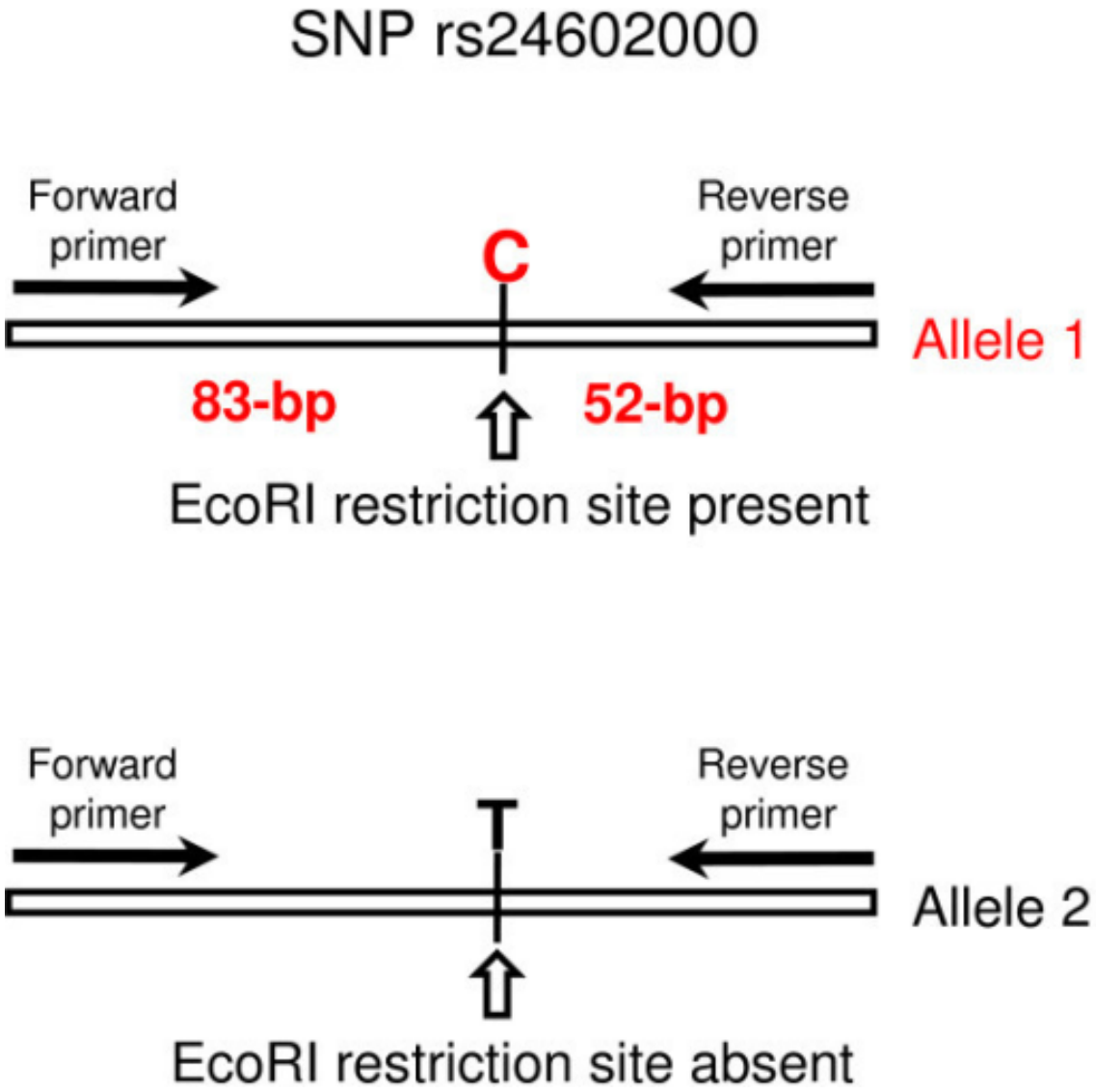
- A lack of neutrality can be a disadvantage if a marker is being used to test whether or not populations are genetically distinct from one another.
  -
- On the other hand, a non-neutral marker can be useful if we are looking for evidence of adaptation.
  - i.e. Colias meadii Populations showed a high level of genetic uniformity **within habitats** but a marked and abrupt shift in allele frequencies between habitats.

RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

- RFLP data are generated using restriction enzymes.



PCR-RFLP

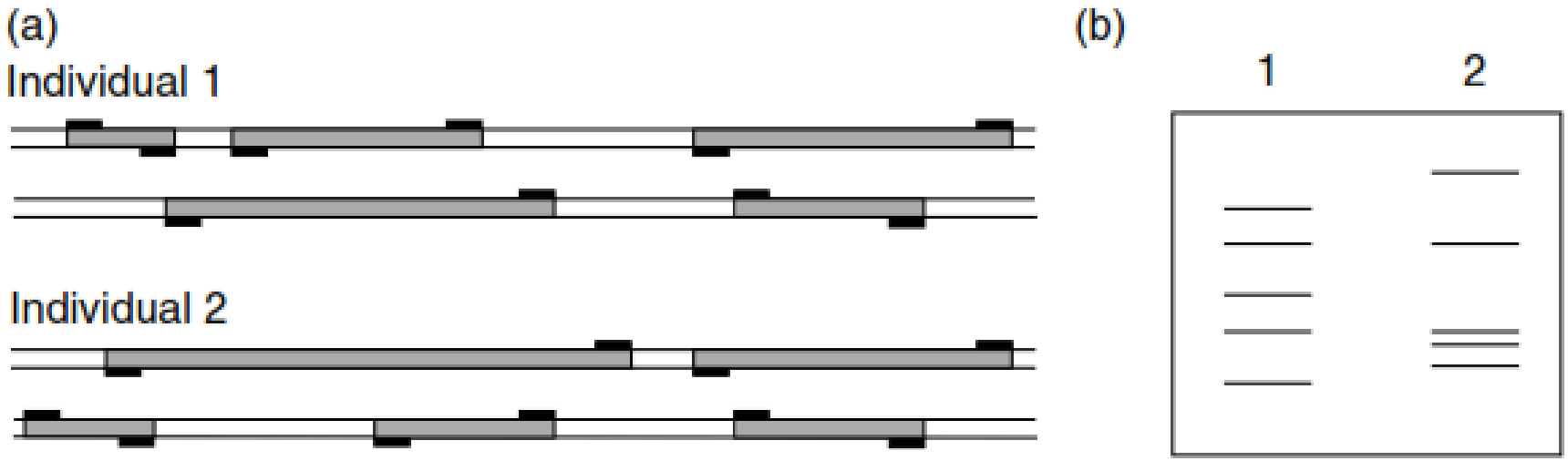

SNP rs24602000
Forward primer
Reverse primer
C
Allele 1
83-bp
52-bp
EcoRI restriction site present
Forward primer
Reverse primer
Allele 2
EcoRI restriction site absent

DOMINANT MARKERS

- Dominant markers are also known as multi-locus markers because they simultaneously generate data from multiple loci. They typically work by using random primers to amplify anonymous regions of the genome, producing a pattern of multiple bands from each individual.

RANDOM AMPLIFIED POLYMORPHIC DNA

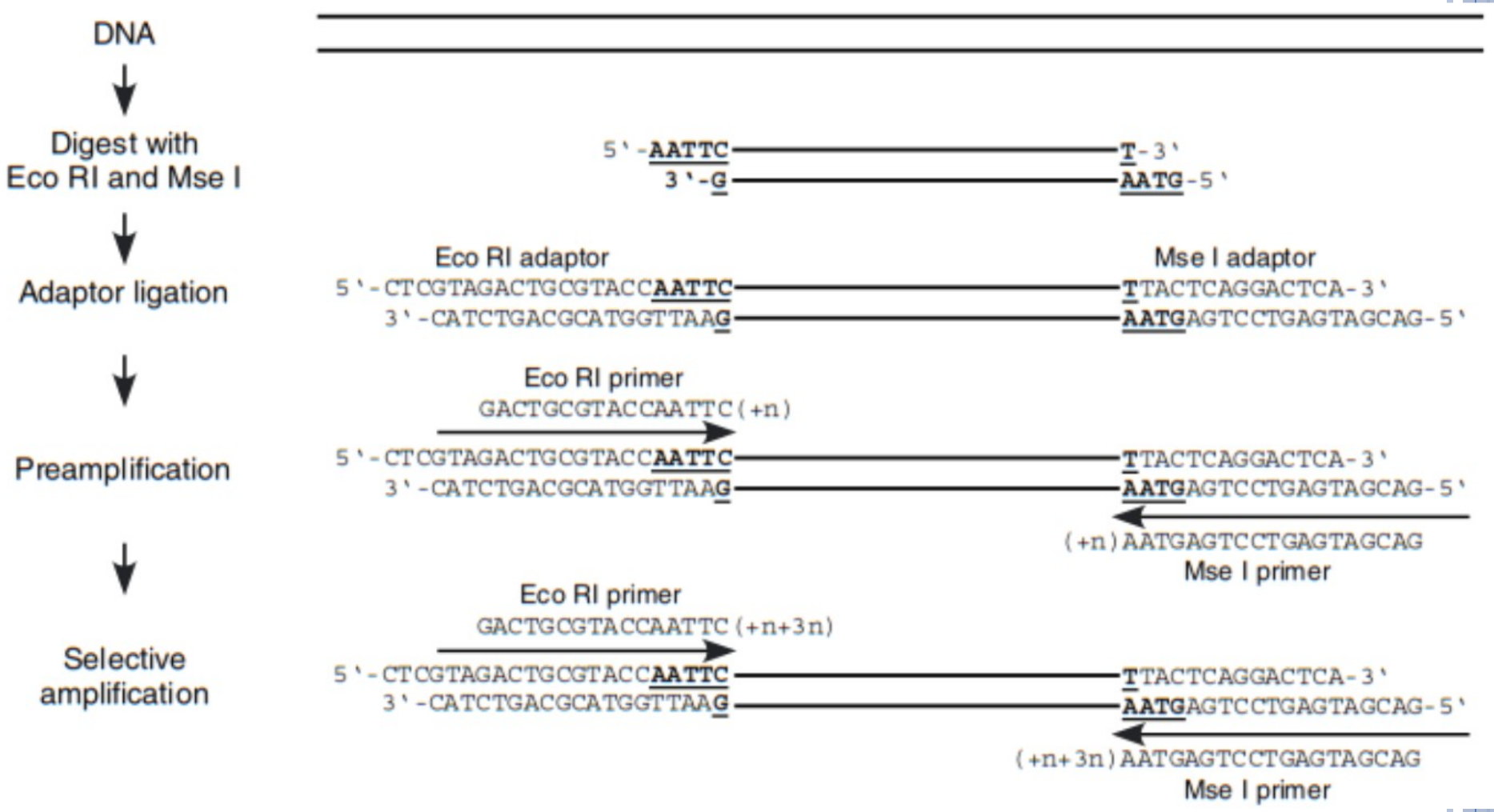
- RAPDs are generated using short (usually 10 bp) random primers in a PCR reaction.



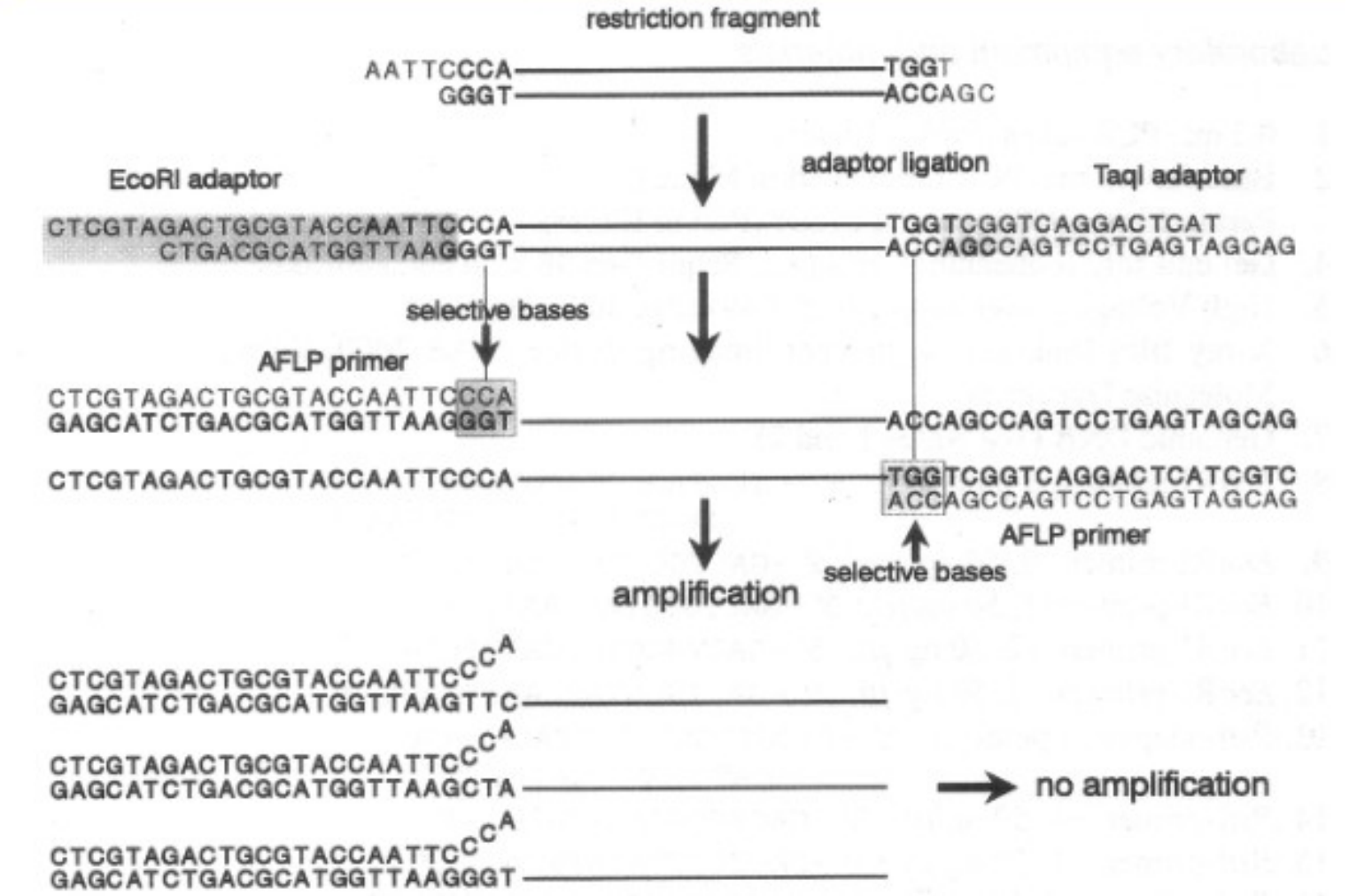
AMPLIFIED FRAGMENT LENGTH POLYMORPHISMS

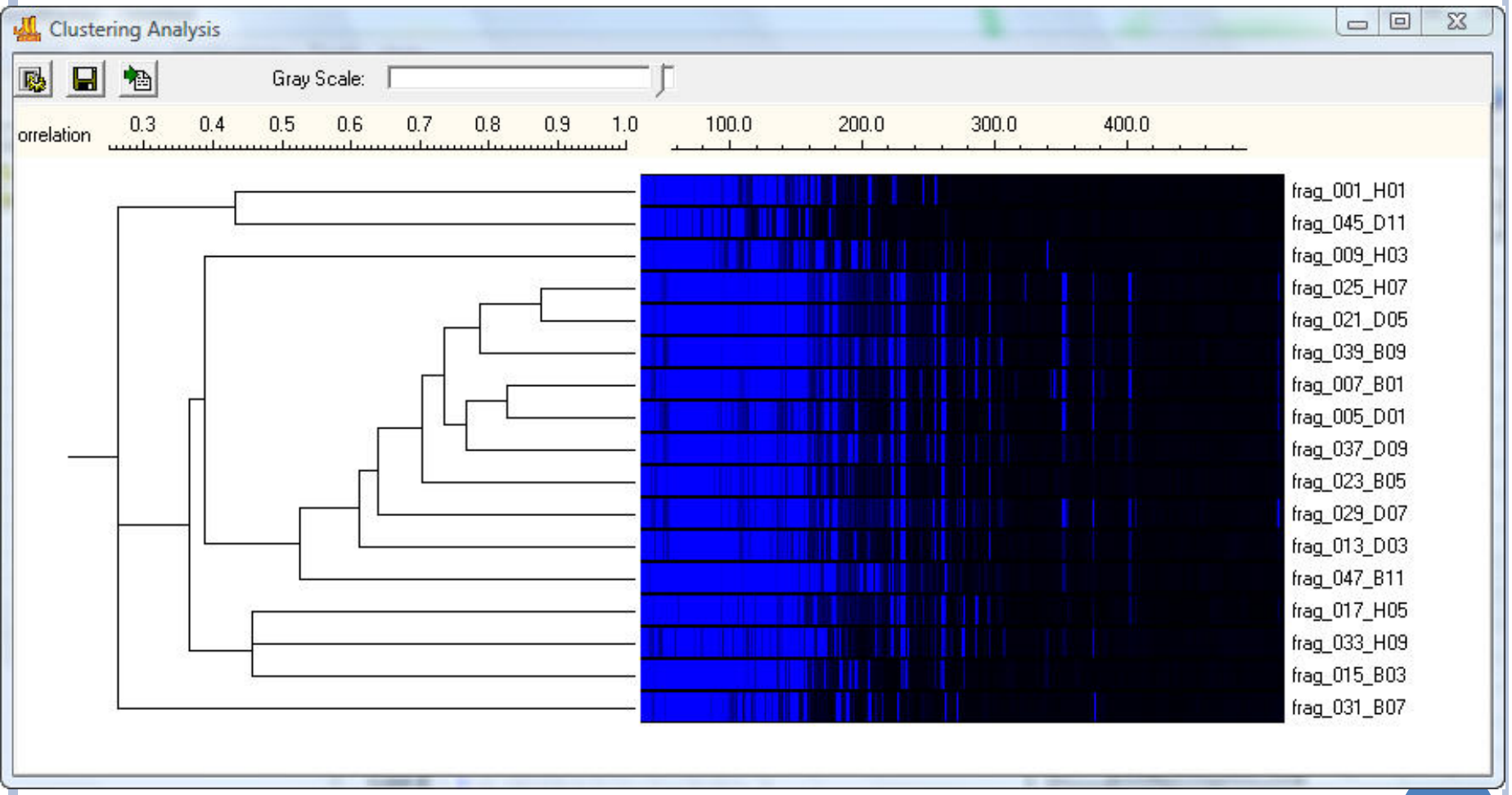
- A more labour intensive, but also more reliable, method than RAPDs for generating PCR-based multi-band profiles is known as amplified fragment length polymorphism (AFLPs)(Voset al., 1995)

AFLP

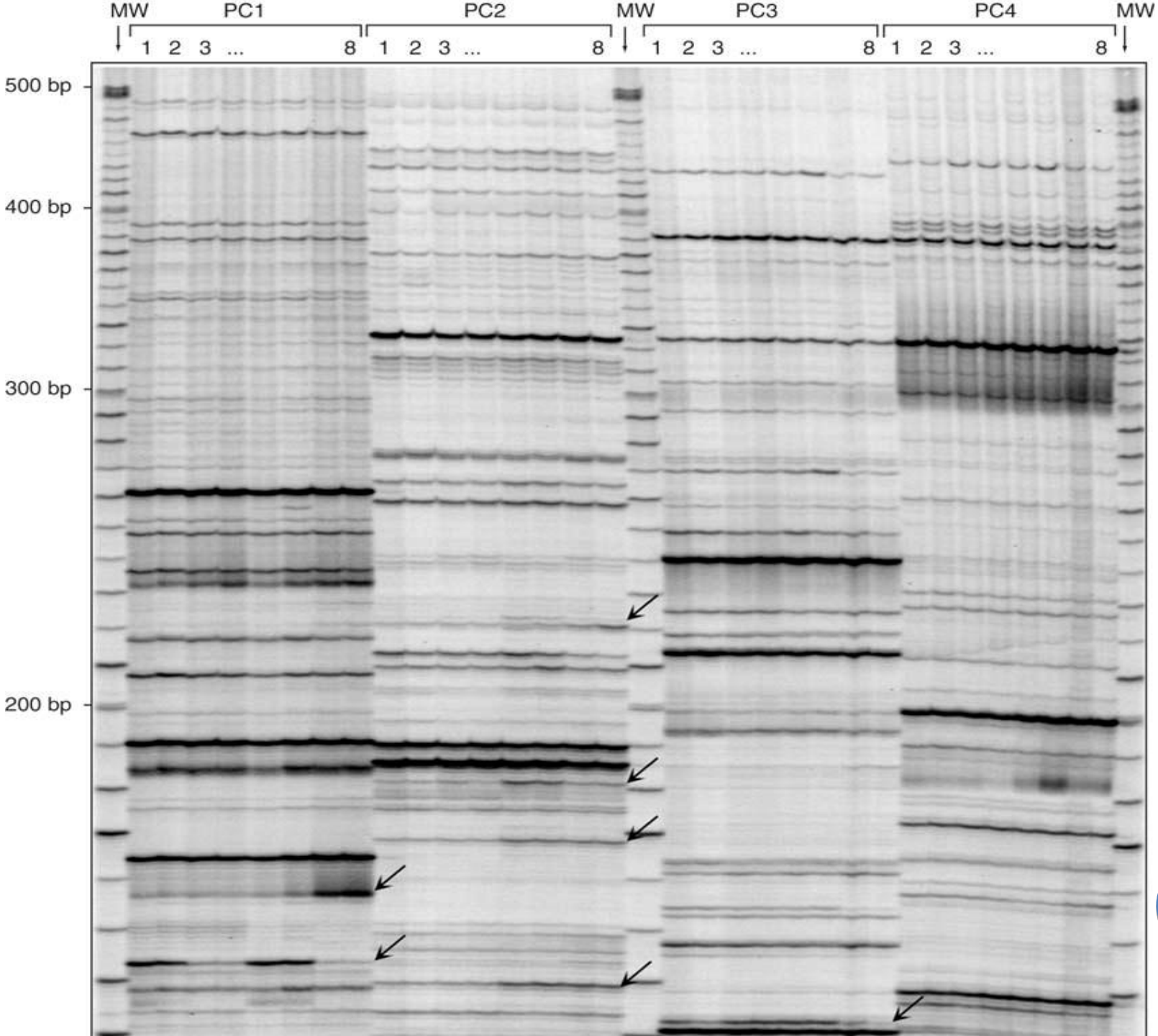

DNA
Digest with Eco RI and Mse I
Adaptor ligation
Preamplification
Selective amplification
5'-AATTC
T-3'
3'-G
AATG-5'
Eco RI adaptor
Mse I adaptor
5'-CTCGTAGACTGCGTACCAATTC
TTACTCAGGACTCA-3'
3'-CATCTGACGCATGGTTAAG
AATGAGTCCTGAGTAGCAG-5'
Eco RI primer
GACTGCGTACCAATTC(+n)
(+n)AATGAGTCCTGAGTAGCAG
Mse I primer
GACTGCGTACCAATTC(+n+3n)
(+n+3n)AATGAGTCCTGAGTAGCAG

SELECTIVE PRIMERS



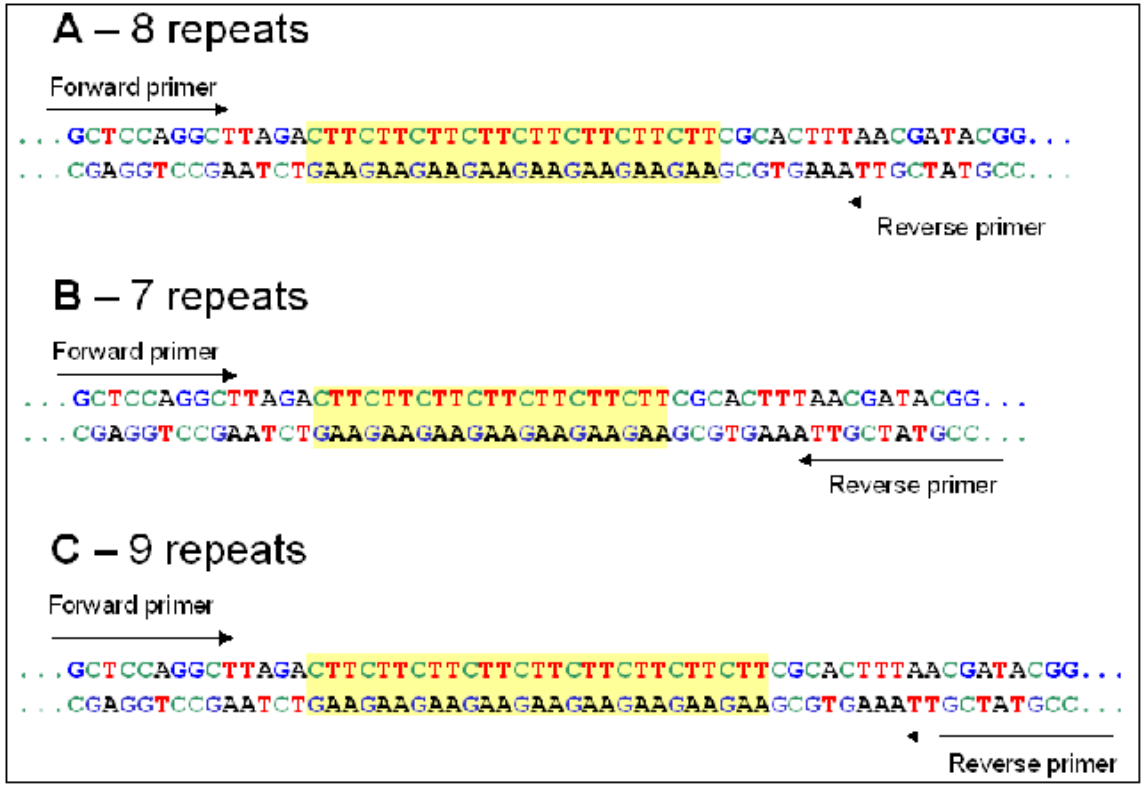

Clustering Analysis
Gray Scale:
orrelation
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
100.0
200.0
300.0
400.0
frag_001_H01
frag_045_D11
frag_009_H03
frag_025_H07
frag_021_D05
frag_039_B09
frag_007_B01
frag_005_D01
frag_037_D09
frag_023_B05
frag_029_D07
frag_013_D03
frag_047_B11
frag_017_H05
frag_033_H09
frag_015_B03
frag_031_B07

- The genetic similarity of individuals and populations can be inferred from the numbers of AFLP bands that they have in common.
- Additional information can be obtained by modifying the standard AFLP method to study **gene expression (cDNA-AFLP) .**

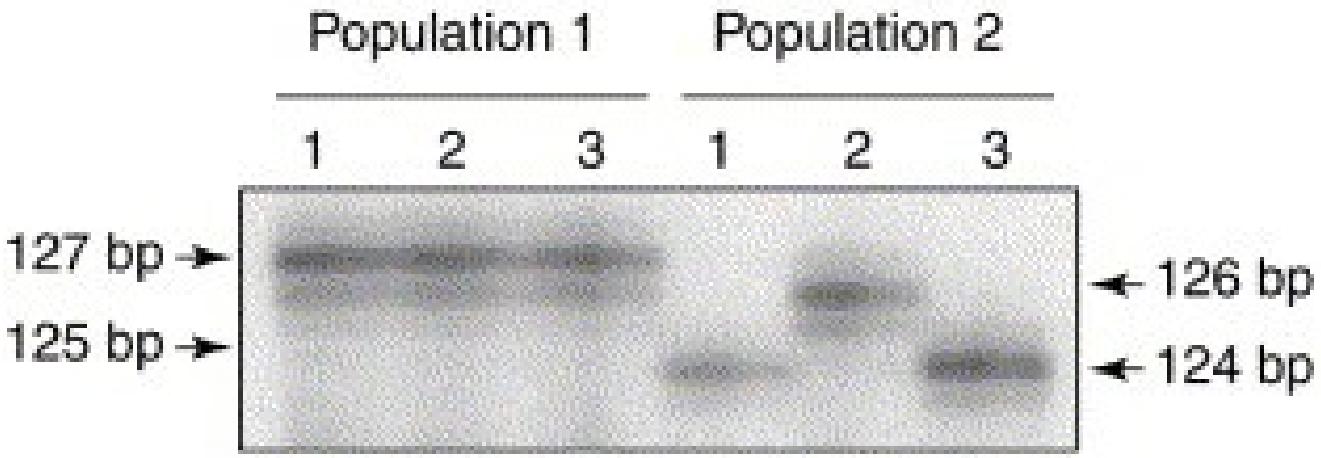

MW
PC1
PC2
MW
PC3
PC4
MW
1 2 3 ... 8
1 2 3 ... 8
1 2 3 ... 8
1 2 3 ... 8
500 bp
400 bp
300 bp
200 bp

MICROSATELLITES

- Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are stretches of DNA that consist of tandem repeats of 1--6 bp.

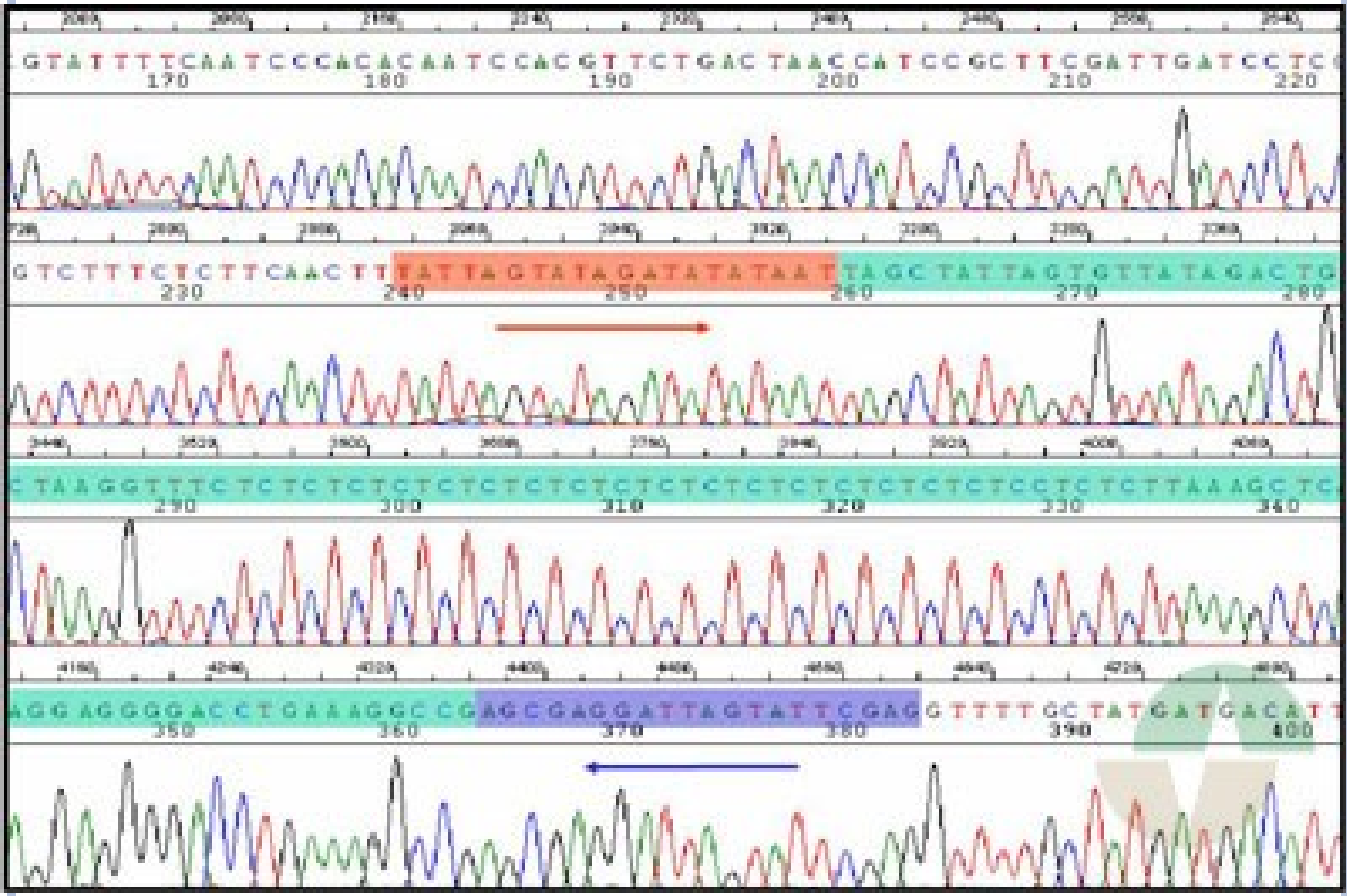


CHLOROPLAST MICROSATELLITES




```
1_1   GACCTCCTTT TTTTTTTTTT ATTTTATATC
1_2   GACCTCCTTT TTTTTTTTTT ATTTTATATC
1_3   GACCTCCTTT TTTTTTTTTT ATTTTATATC
2_1   GACCTCCTTT TTTTTTT... ATTTTATATC
2_2   GACCTCCTTT TTTTTTTTT. ATTTTATATC
2_3   GACCTCCTTT TTTTTTT... ATTTTATATC
```

TRENDS in Ecology & Evolution


GTATTTTCAATCCCACACAATCCACGTTCTGACTAACCATCCGCTTCGATTGATCCTC
170
180
190
200
210
220
GTCTTTCTCTTCAACTTTATTAGTATAGATATATAATTAGCTATTAGTGTTATAGACTG
230
240
250
260
270
280
CTAAGGTTTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCCTCTCTTAAAGCTC
290
300
310
320
330
340
AGGAGGGGACCTGAAAGGCCGAGCGAGGATTAGTATTCGAGGTTTTGCTATGATGACATT
350
360
370
380
390
400

MICROSATELLITES MUTATION RATE

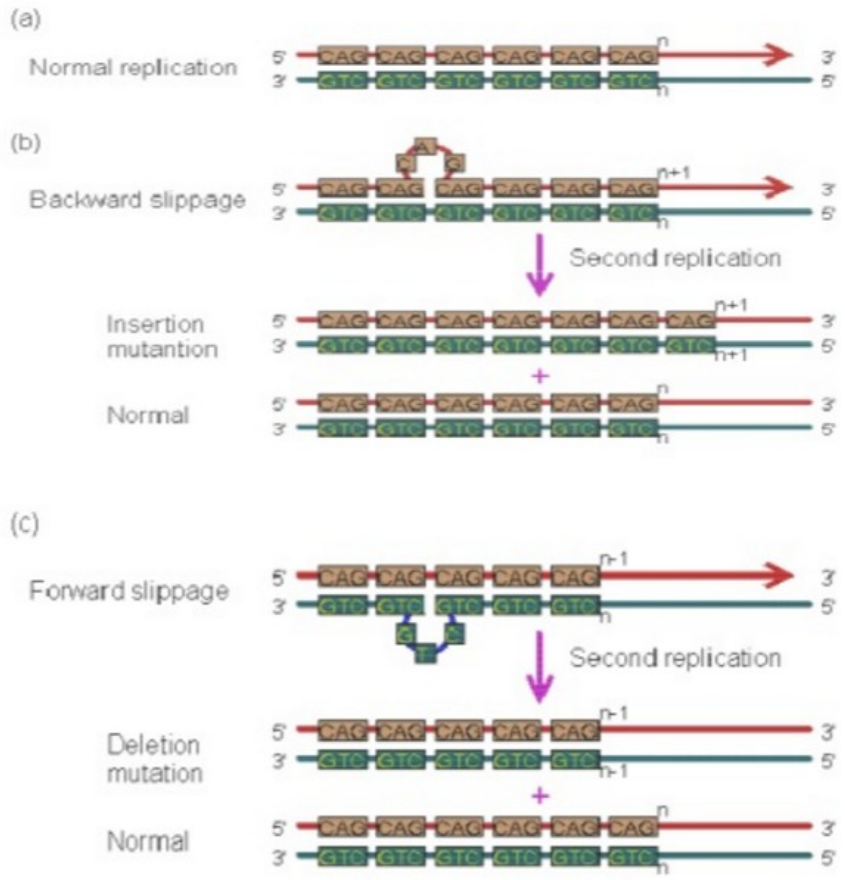
- Microsatellites mutate much more rapidly than most other types of sequences, with estimated mutation rates of around 10^{-4}—10^{-5} events per locus per replication in year and around 10^{-3}—10^{-4} in mice.
- This is substantially higher than the estimated overall point mutation rate.

How do microsatellites mutate?

- Replication Slippage
- Unequal crossing-over during meiosis

REPLICATION SLIPPAGE

When the DNA replicates, the polymerase loses track of its place, and either leaves out repeat units or adds too many repeat units.



"Polymerase slippage" or "slipped-strand mispairing."

A commonly observed replication error is the **replication slippage**, which occurs at the repetitive sequences when the new strand mispairs with the template strand. The microsatellite polymorphism is mainly caused by the replication slippage. If the mutation occurs in a coding region, it could produce abnormal proteins, leading to diseases.

MICROSATELLITES MUTATION MODELS

- **stepwise mutation model** (SMM; Kimura and Ohta, 1978).
- **infinite alleles model** (IAM; Kimura and Crow, 1964)

MICROSATELLITES MUTATION MODELS

1. Stepwise Mutation Model (SMM)

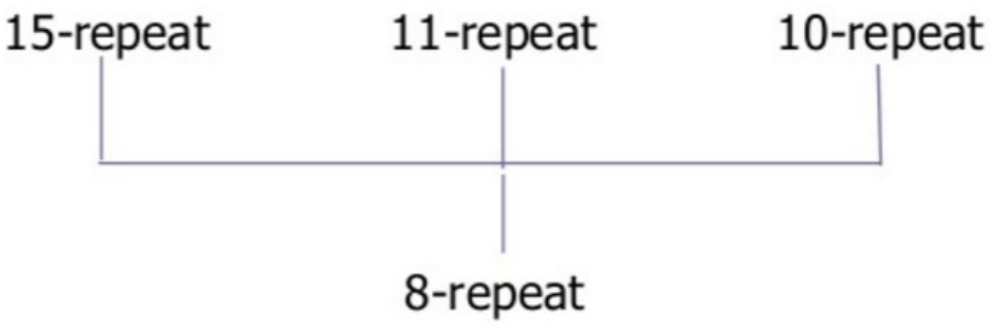
This model holds that when microsatellites mutate, they only gain or lose one repeat. This implies that two alleles that differ by one repeat are more closely related (have a more recent common ancestor) than alleles that differ by many repeats. In other words, size matters when doing statistical tests of population substructuring. The SMM is generally the preferred model when calculating relatedness between individuals and population substructuring, although there is the problem of homoplasy.

MICROSATELLITES MUTATION MODELS

2. Infinite Alleles Model (IAM)

Each mutation can create any new allele randomly. A 15-repeat allele could be just as closely related to a 10-repeat allele as a 11-repeat allele. All that matters is that they are different alleles. In other words, size isn't important.

A 15-repeat allele could be just as closely related to a 10-repeat allele as a 11-repeat allele.



MICROSATELLITES MUTATION RATE

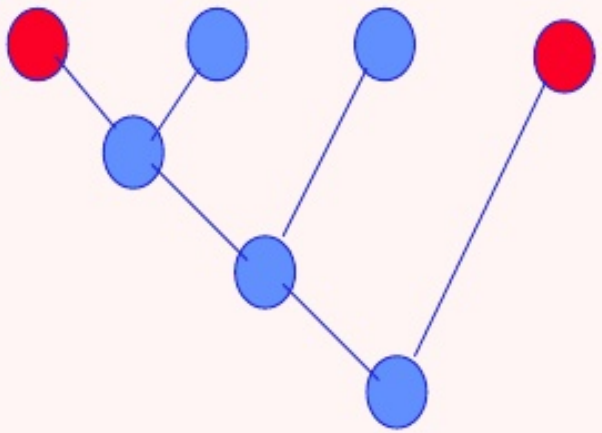
- the high mutation rates of microsatellites mean that there are often multiple alleles at each locus, and this high level of polymorphism makes them suitable for inferring relatively recent population genetic events.

HOMOPLASY

- Microsatellite data are not particularly useful for inferring evolutionary events that occurred in the relatively distant past. Their rapid rate of mutation and their tendency to either increase or decrease in size means that size **homoplasy** may often occur.

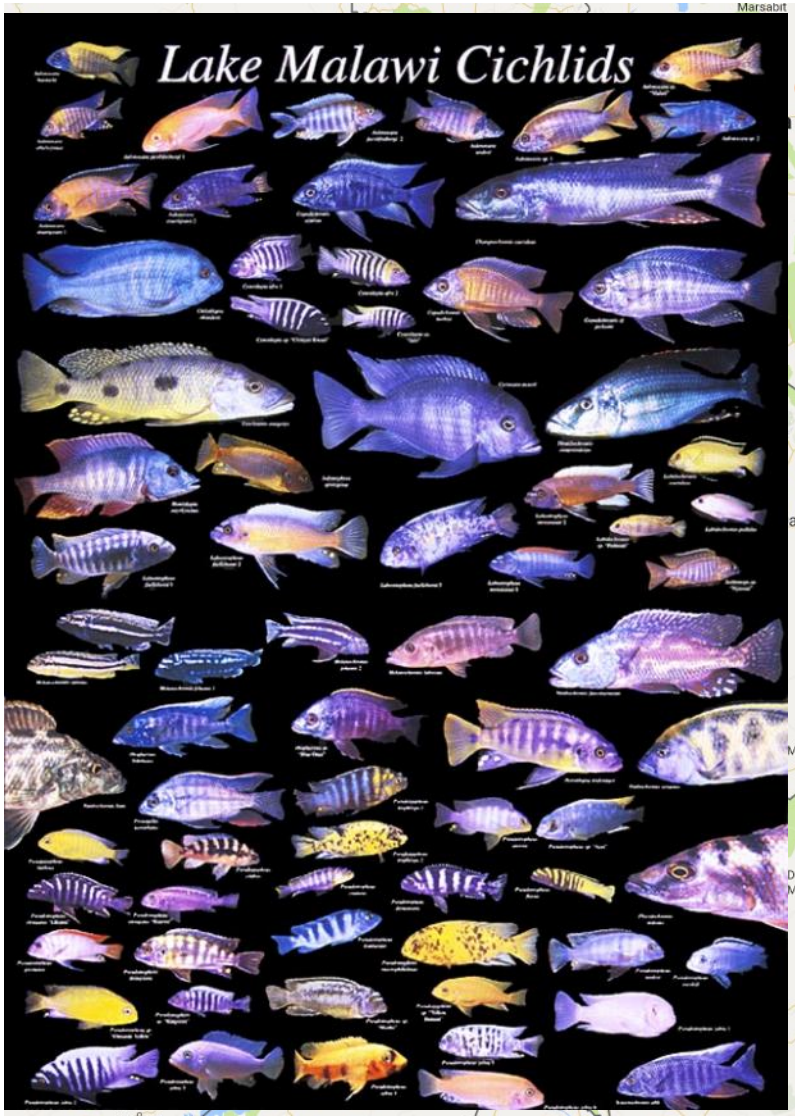
Homoplasy:

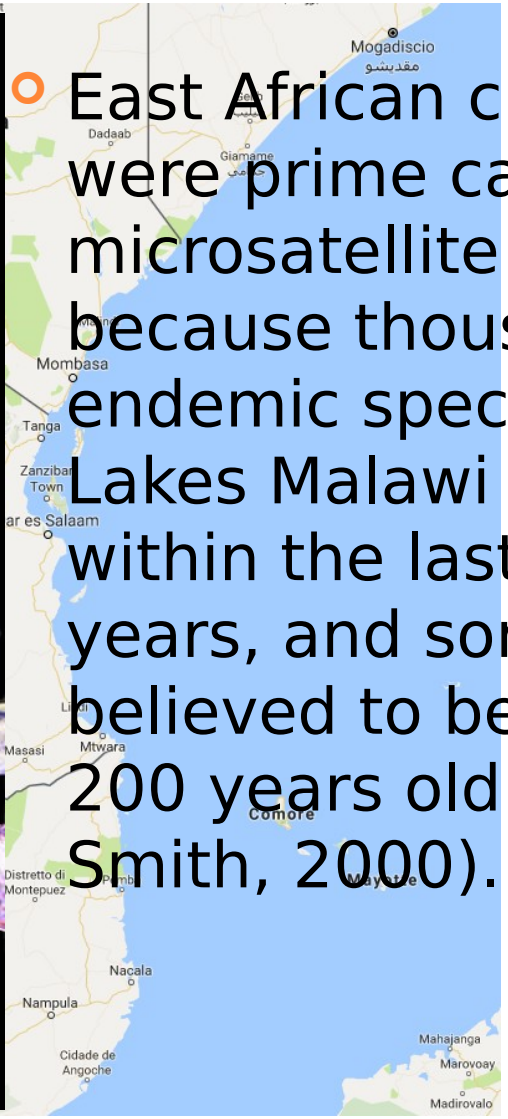
Homoplasy is a poor indicator of evolutionary relationships because the similarity does not reflect shared ancestry.



It is sometimes useful to distinguish between different types of homoplasy ....
Convergence, Parallel substitution and Reversals (Secondary Loss)

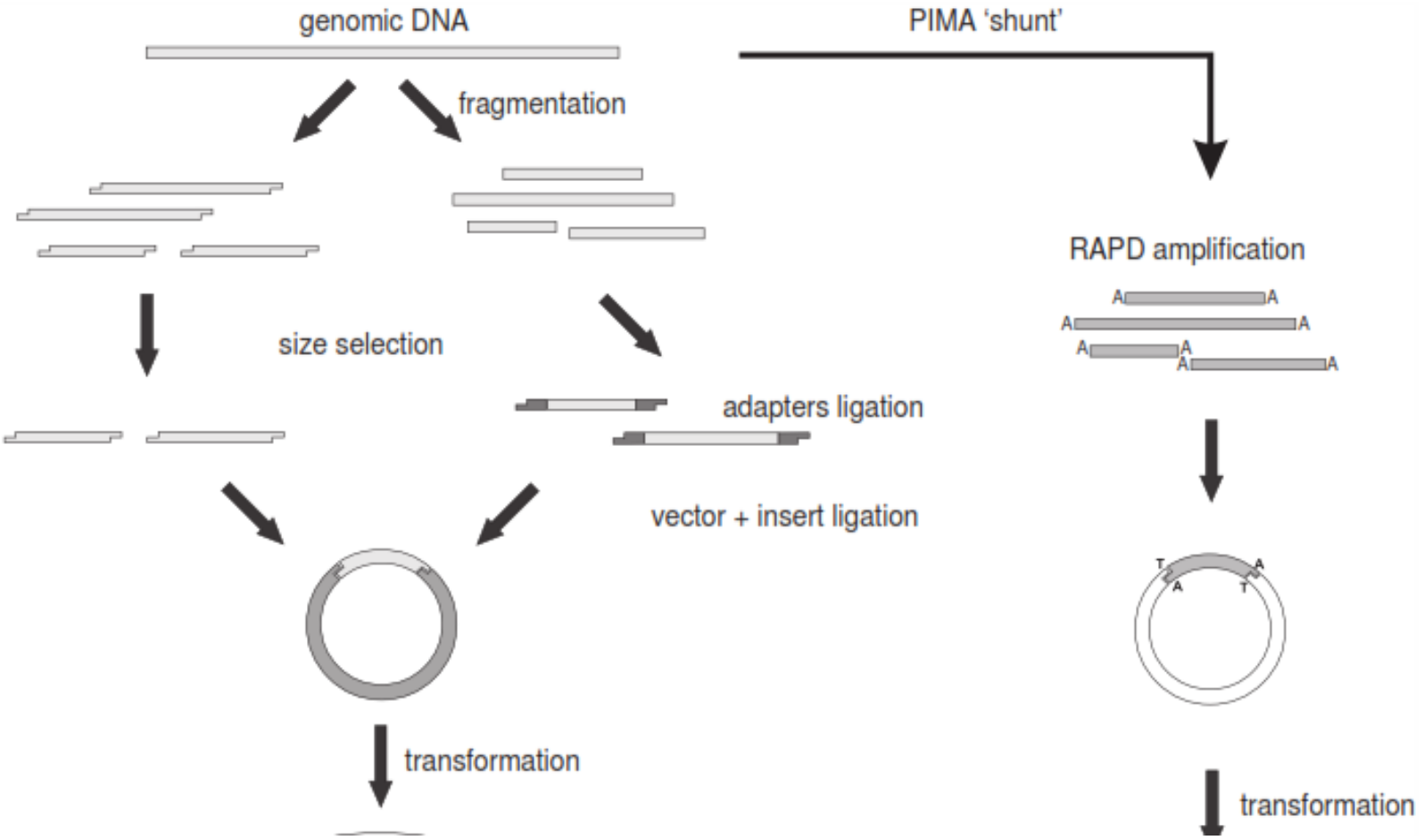
…BUT GOOD FOR RECENT EVOLUTIONARY EVENTS.

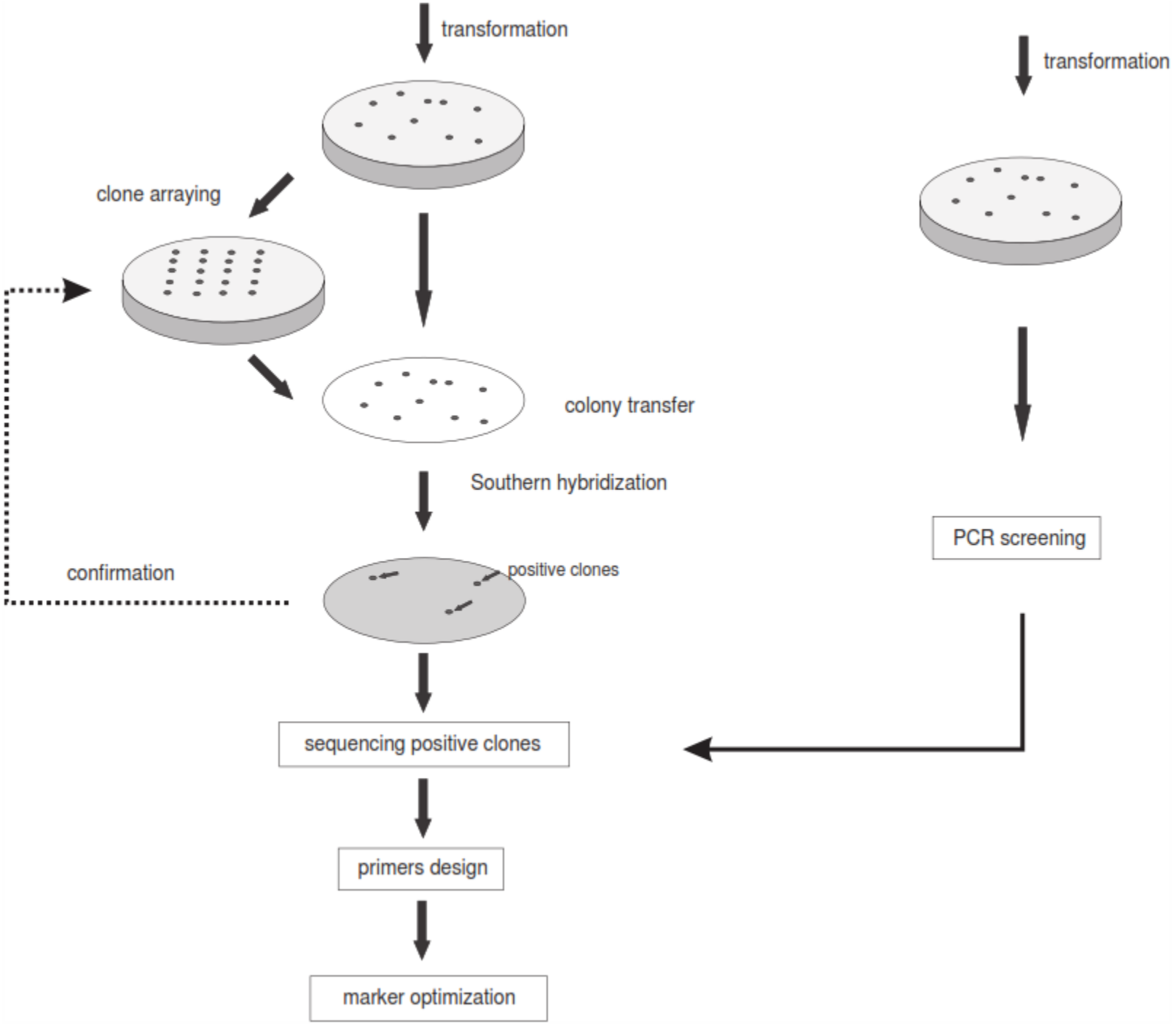




- East African cichlid fishes were prime candidates for microsatellite analysis, because thousands of endemic species evolved in Lakes Malawi and Victoria within the last 700 000 years, and some species are believed to be only around 200 years old (Kornfield and Smith, 2000).

SSR ISOLATION




transformation
clone arraying
colony transfer
Southern hybridization
confirmation
positive clones
sequencing positive clones
primers design
marker optimization
transformation
PCR screening

DNA SEQUENCES

- DNA sequences can be obtained from fragments of DNA that have been amplified by PCR.
- Although sequence data can be extremely informative, obtaining these data is quite expensive (although decreasingly so) and time-consuming. Development time will be longer if a number of sequences need to be screened before appropriately variable regions are identified.

- Furthermore, many studies benefit by having data from more than one genetic region, which further adds to the time and expense. Recently, however, a relatively new method known as single nucleotide polymorphisms (SNPs) has been gaining in popularity because it is specifically designed to target variable DNA bases in multiple loci.

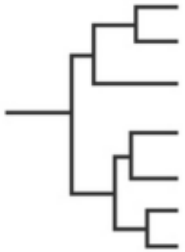
SINGLE NUCLEOTIDE POLYMORPHISMS

- Single nucleotide polymorphisms (SNPs) refer to single base pair positions along a DNA sequence that vary between individuals.
- Most SNPs have only two alternative states (i.e. each individual has one of two possible nucleotides at a given SNP locus) and are therefore referred to as **biallelic markers**.

RAD SEQUENCING DISCOVERY AND GENOTYPING

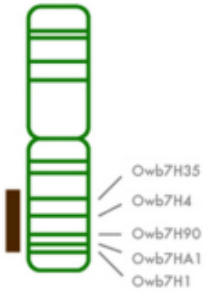
- Restriction site Associated DNA Sequencing (RAD-Seq) is a fractional genome sequencing strategy, designed to interrogate anywhere from 0.1 to 10% of a selected genome.

- By digesting the genome with a restriction nuclease and attaching a series of adapters to the resulting DNA fragments, large numbers of genetic variations such as SNPs can be readily identified from analysis of next generation DNA sequence data.

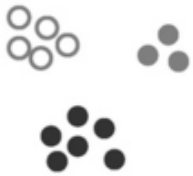


PHYLOGENETIC ANALYSIS

POPULATION STRUCTURE

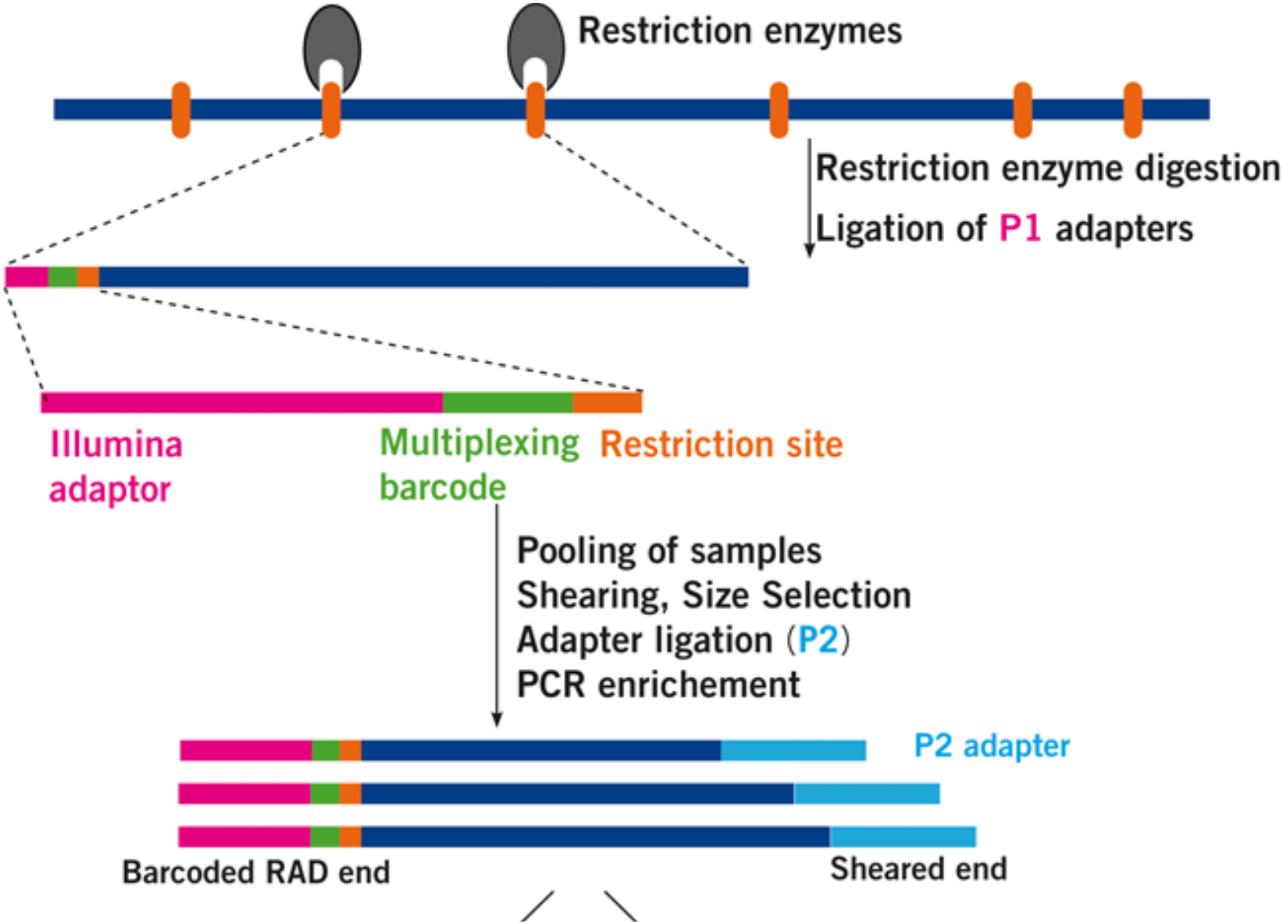


LINKAGE AND QTL MAPPING

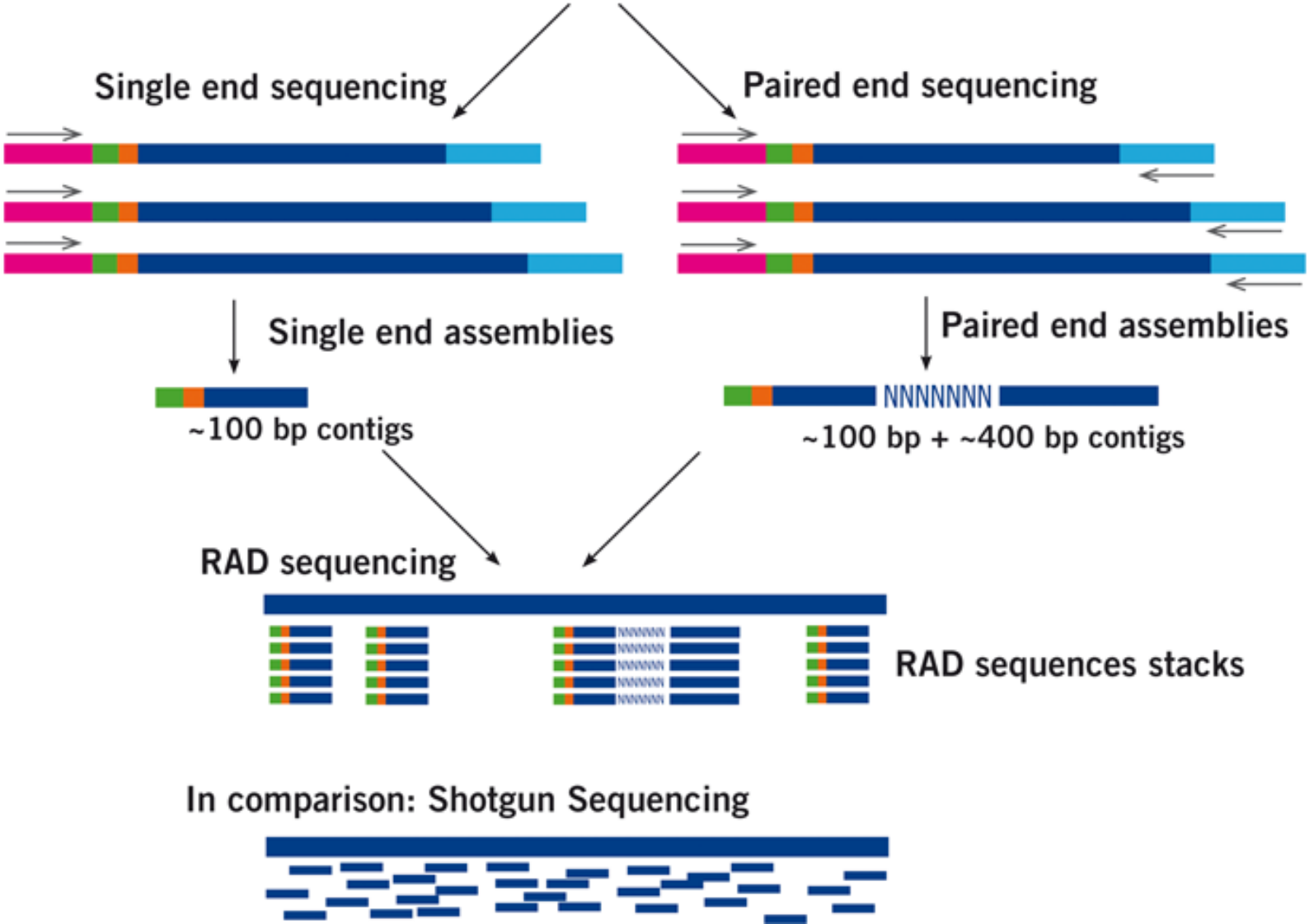


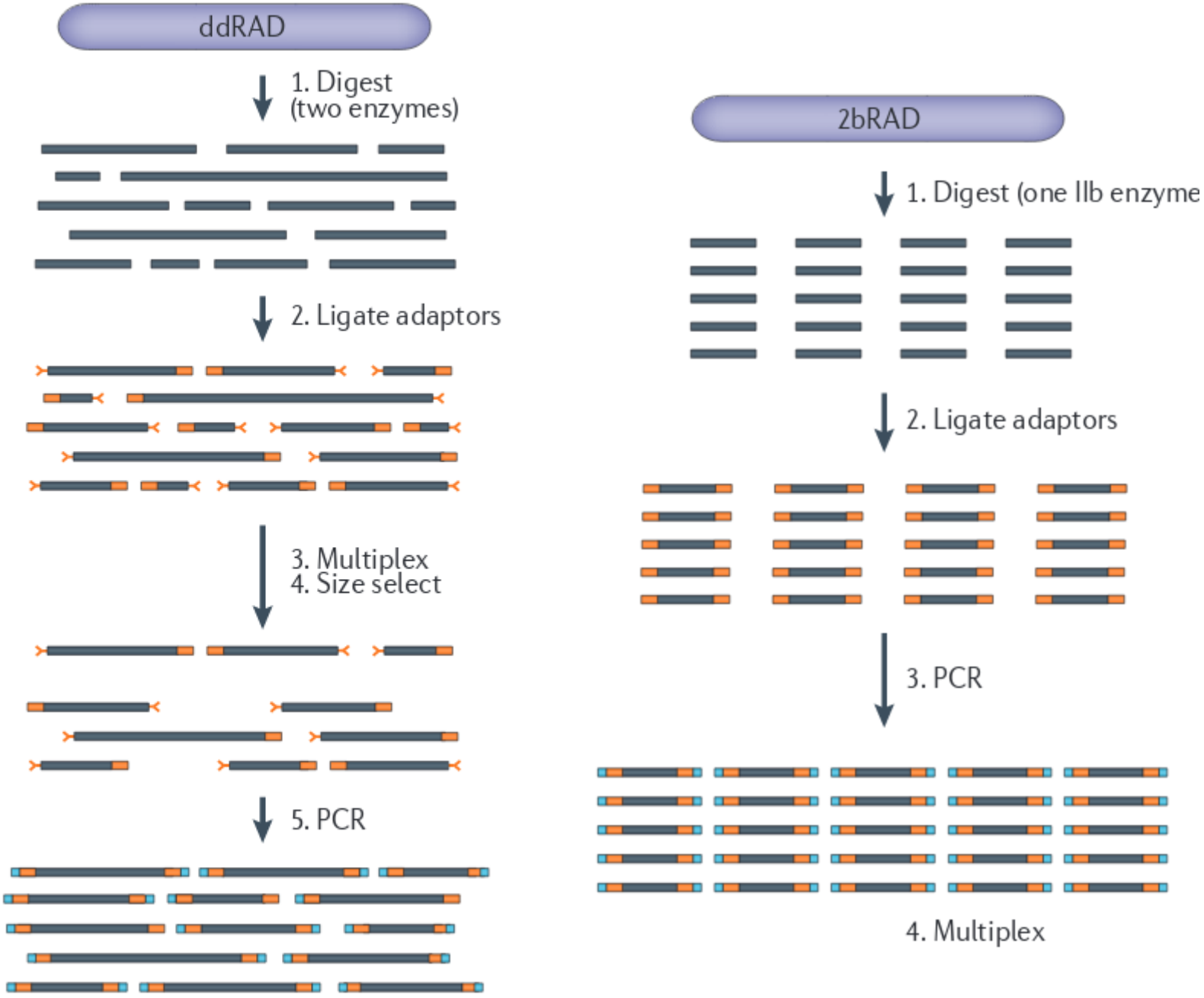
ASSOCIATION MAPPING

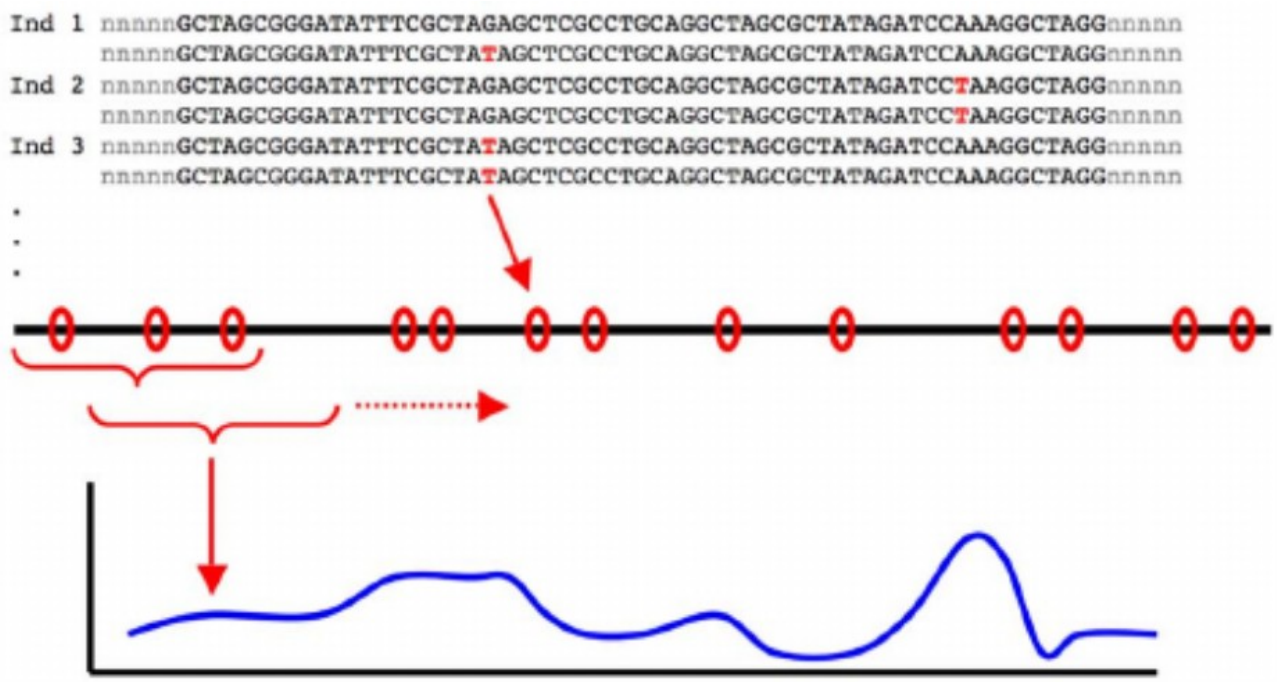
RADSEQ


Restriction enzymes
Restriction enzyme digestion
Ligation of P1 adapters
Illumina adaptor
Multiplexing barcode
Restriction site
Pooling of samples
Shearing, Size Selection
Adapter ligation (P2)
PCR enrichement
P2 adapter
Barcoded RAD end
Sheared end

RADSEQ


Single end sequencing
Paired end sequencing
Single end assemblies
Paired end assemblies
~100 bp contigs
NNNNNNN
~100 bp + ~400 bp contigs
RAD sequencing
RAD sequences stacks
In comparison: Shotgun Sequencing


ddRAD
1. Digest
(two enzymes)
2. Ligate adaptors
3. Multiplex
4. Size select
5. PCR
2bRAD
1. Digest (one IIb enzyme
2. Ligate adaptors
3. PCR
4. Multiplex


Ind 1 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
Ind 2 nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTAGAGCTCGCCTGCAGGCTAGCGCTATAGATCCTAAGGCTAGGnnnnn
Ind 3 nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn
nnnnnGCTAGCGGGATATTTCGCTATAGCTCGCCTGCAGGCTAGCGCTATAGATCCAAAGGCTAGGnnnnn

Genomics of adaptation