CENTER OF
PLANT SCIENCES

Sant'Anna
Scuola Universitaria Superiore Pisa

Advanced Genomics

Navigating diversity

How small can a genome be?

- Pioneering work by Craig Venter
- What is the minimum number of genes to support life?
- *Mycoplasma genitalium*
  - Parasite bacteria
  - 580Kb genome
  - A few hundreds genes

Essential genes of a minimal bacterium

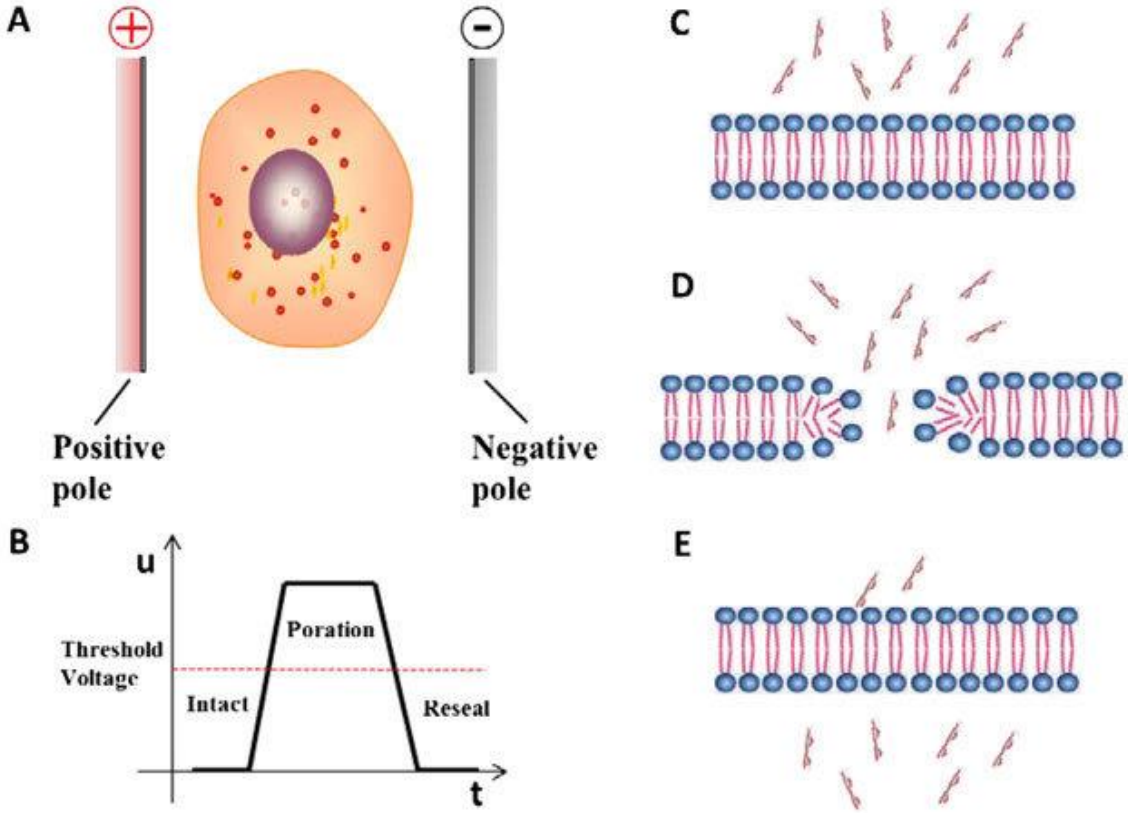
John I. Glass, Nacyra Assad-Garcia, Nina Alperovich, Shibu Yooseph, Matthew R. Lewis, Mahir Maruf, Clyde A. Hutchison III, Hamilton O. Smith*, and J. Craig Venter

Synthetic Biology Group, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850

Contributed by Hamilton O. Smith, November 18, 2005

Mycoplasma genitalium **has the smallest genome of any organism that can be grown in pure culture. It has a minimal metabolism and little genomic redundancy. Consequently, its genome is expected to be a close approximation to the minimal set of genes needed to sustain bacterial life. Using global transposon mutagenesis, we isolated and characterized gene disruption mutants for 100 different nonessential protein-coding genes. None of the 43 RNA-coding genes were disrupted. Herein, we identify 382 of the 482 *M. genitalium* protein-coding genes as essential, plus five sets of disrupted genes that encode proteins with potentially redundant essential functions, such as phosphate transport. Genes encoding proteins of unknown function constitute 28% of the essential protein-coding genes set. Disruption of some genes accelerated *M. genitalium* growth.**

- Transformation of *M. genitalium* with Tn4001 by Electroporation
- Amplification of Isolated Colonies for DNA Extraction
- Sequencing to retrieve Tn4001 insertions



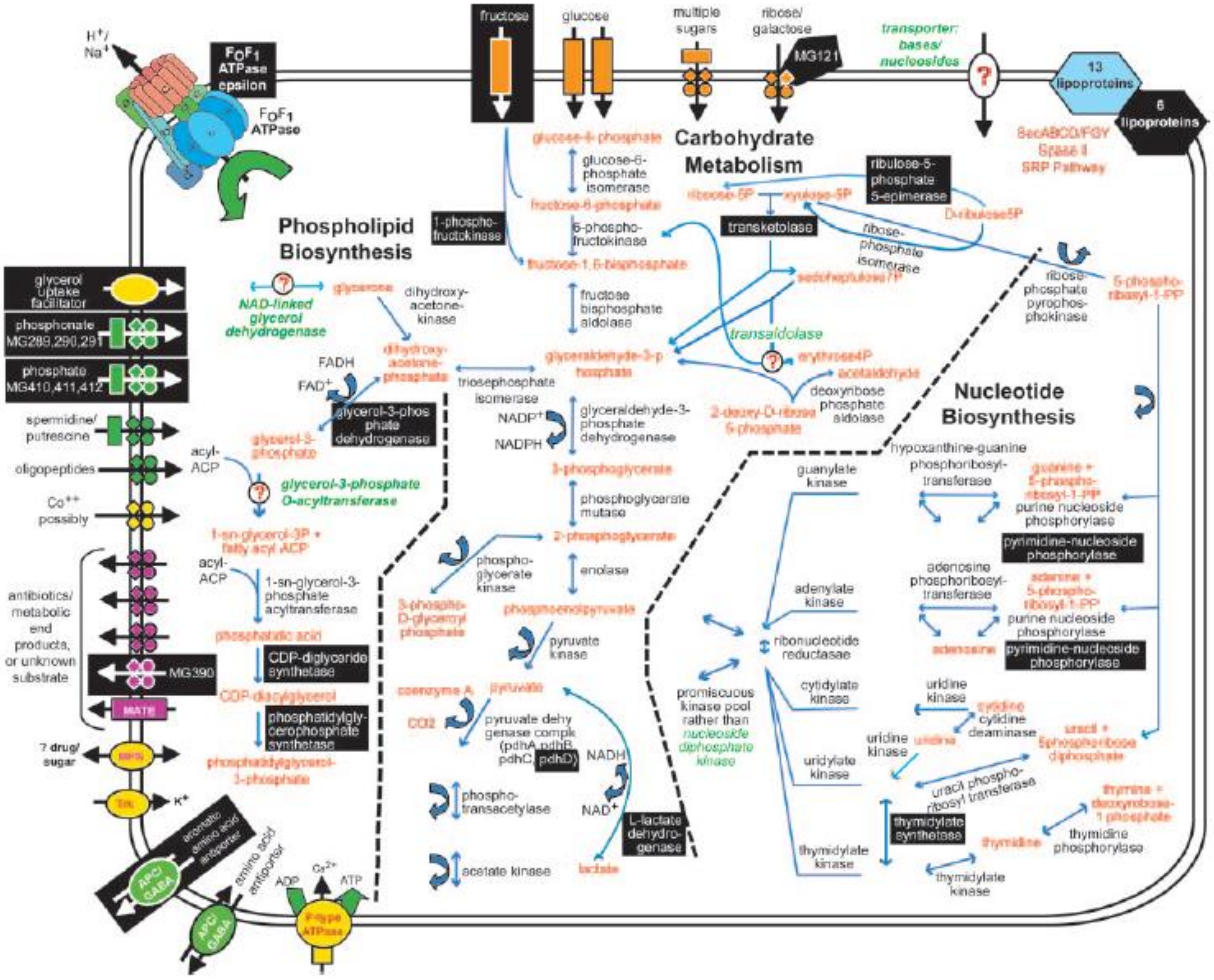
Short Communication

Tn*4001*: A Gentamicin and Kanamycin Resistance Transposon in *Staphylococcus aureus*

Bruce R. Lyon, John W. May, and Ronald A. Skurray
Department of Microbiology, Monash University, Clayton, Victoria, Australia 3168

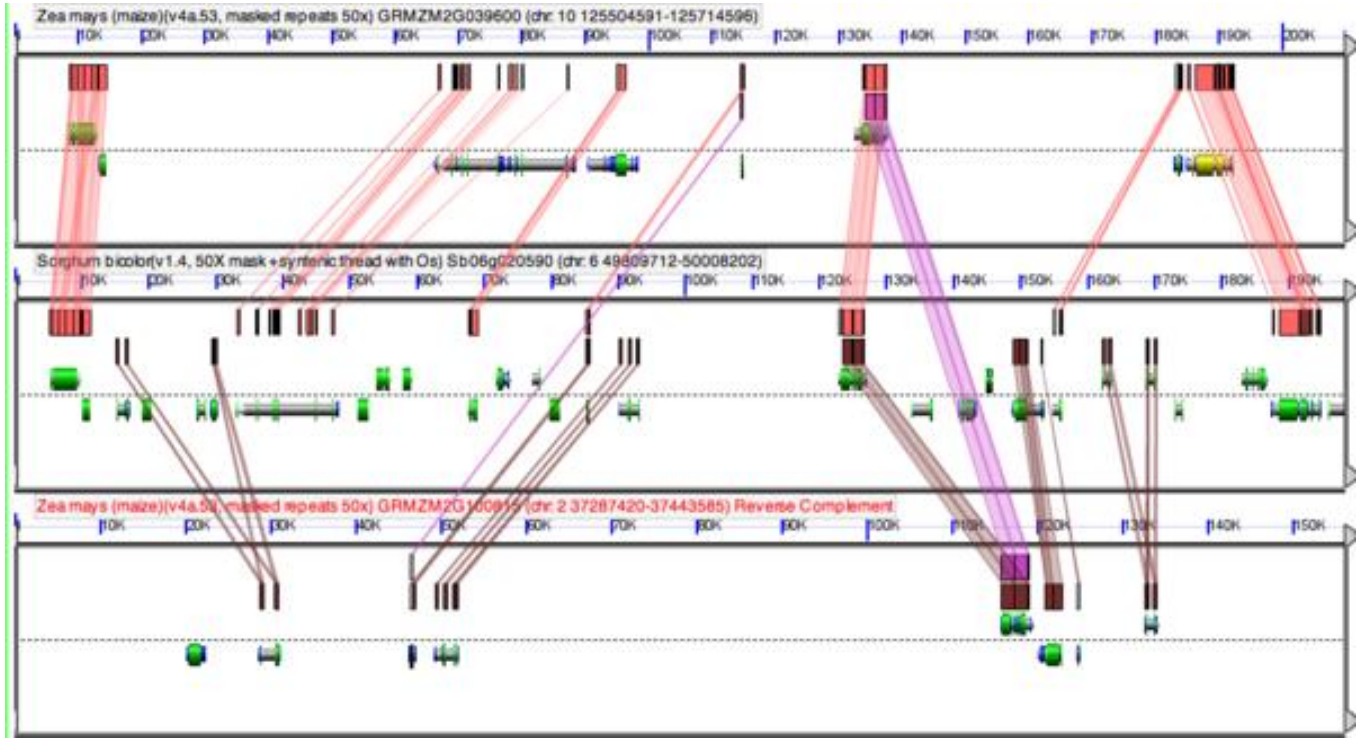
Summary. We describe a 4.5 kilobase transposon, Tn*4001*, which mediates resistance to gentamicin, tobramycin and kanamycin in *Staphylococcus aureus*. Originally detected in plasmid pSK1, Tn*4001* was shown to undergo *rec*-independent transposition to the chromosome from this plasmid and from an inserted derivative of the plasmid pII147. Heteroduplexes between plasmids with and without Tn*4001* demonstrated a characteristic stem and loop structure with inverted repeats of approx. 1.3 kilobases.

- Whichever insertion you can retrieve, it is not an essential gene
- The minimum number of genes in *Mg* is 382 (the number varies by species)



fructose
glucose
multiple sugars
ribose/ galactose
MG121
transporter: bases/ nucleosides
13 lipoproteins
6 lipoproteins
SecABCD/FGY
Spase II
SRP Pathway
FoF1 ATPase epsilon
FoF1 ATPase
Carbohydrate Metabolism
Phospholipid Biosynthesis
Nucleotide Biosynthesis
glucose-6-phosphate
glucose-6-phosphate isomerase
fructose-6-phosphate
1-phospho-fructokinase
6-phospho-fructokinase
fructose-1,6-bisphosphate
fructose bisphosphate aldolase
glyceraldehyde-3-p hosphate
triosephosphate isomerase
glyceraldehyde-3-phosphate dehydrogenase
3-phosphoglycerate
phosphoglycerate mutase
2-phosphoglycerate
enolase
phosphoenolpyruvate
pyruvate kinase
pyruvate
coenzyme A
CO2
pyruvate dehy genase compk (pdhA, pdhB, pdhC, pdhD)
NADH
NAD+
L-lactate dehydro-genase
lactate
phospho-transacetylase
acetate kinase
phospho-glycerate kinase
3-phospho-D-glyceroyl phosphate
NADP+
NADPH
ribulose-5-phosphate 5-epimerase
ribose-5P
xylulose-5P
D-ribulose5P
transketolase
ribose-phosphate isomerase
sedoheptulose7P
transaldolase
erythrose4P
acetaldehyde
deoxyribose phosphate aldolase
2-deoxy-D-ribose 5-phosphate
ribose-phosphate pyrophos-phokinase
5-phospho-ribosyl-1-PP
glycerol uptake facilitator
phosphonate MG289,290,291
phosphate MG410,411,412
spermidine/ putrescine
oligopeptides
Co++ possibly
antibiotics/ metabolic end products, or unknown substrate
MG390
MATE
? drug/ sugar
K+
aromatic amino acid antiporter
amino acid antiporter
APC/ GABA
ADP
Ca2+
ATP
P-type ATPase
NAD-linked glycerol dehydrogenase
glycerone
dihydroxy-acetone-kinase
dihydroxy-acetone-phosphate
FADH
FAD+
glycerol-3-phosphate dehydrogenase
glycerol-3-phosphate
acyl-ACP
glycerol-3-phosphate O-acyltransferase
1-sn-glycerol-3P + fatty acyl ACP
1-sn-glycerol-3-phosphate acyltransferase
phosphatidic acid
CDP-diglyceride synthetase
CDP-diacylglycerol
phosphatidylglycerophosphate synthetase
phosphatidylglycerol-3-phosphate
hypoxanthine-guanine phosphoribosyl-transferase
guanine + 5-phospho-ribosyl-1-PP
purine nucleoside phosphorylase
pyrimidine-nucleoside phosphorylase
guanylate kinase
adenosine phosphoribosyl-transferase
adenine + 5-phospho-ribosyl-1-PP
adenylate kinase
adenosine
ribonucleotide reductase
promiscuous kinase pool rather than nucleoside diphosphate kinase
cytidylate kinase
uridine kinase
cytidine
cytidine deaminase
uridine
uracil + 5phosphoribose diphosphate
uracil phospho-ribosyl transferase
uridylate kinase
thymidylate synthetase
thymine + deoxyribose-1 phosphate
thymidine
thymidine phosphorylase
thymidylate kinase

Different genomes are characterized by different gene content (ref. pangenomes) and different non-coding elements (ref. TEs)

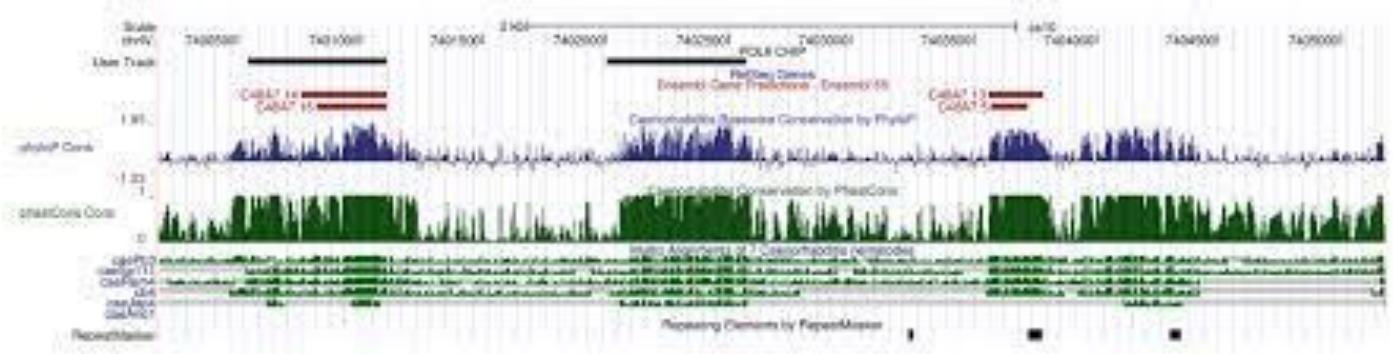


Genomes at a glimpse

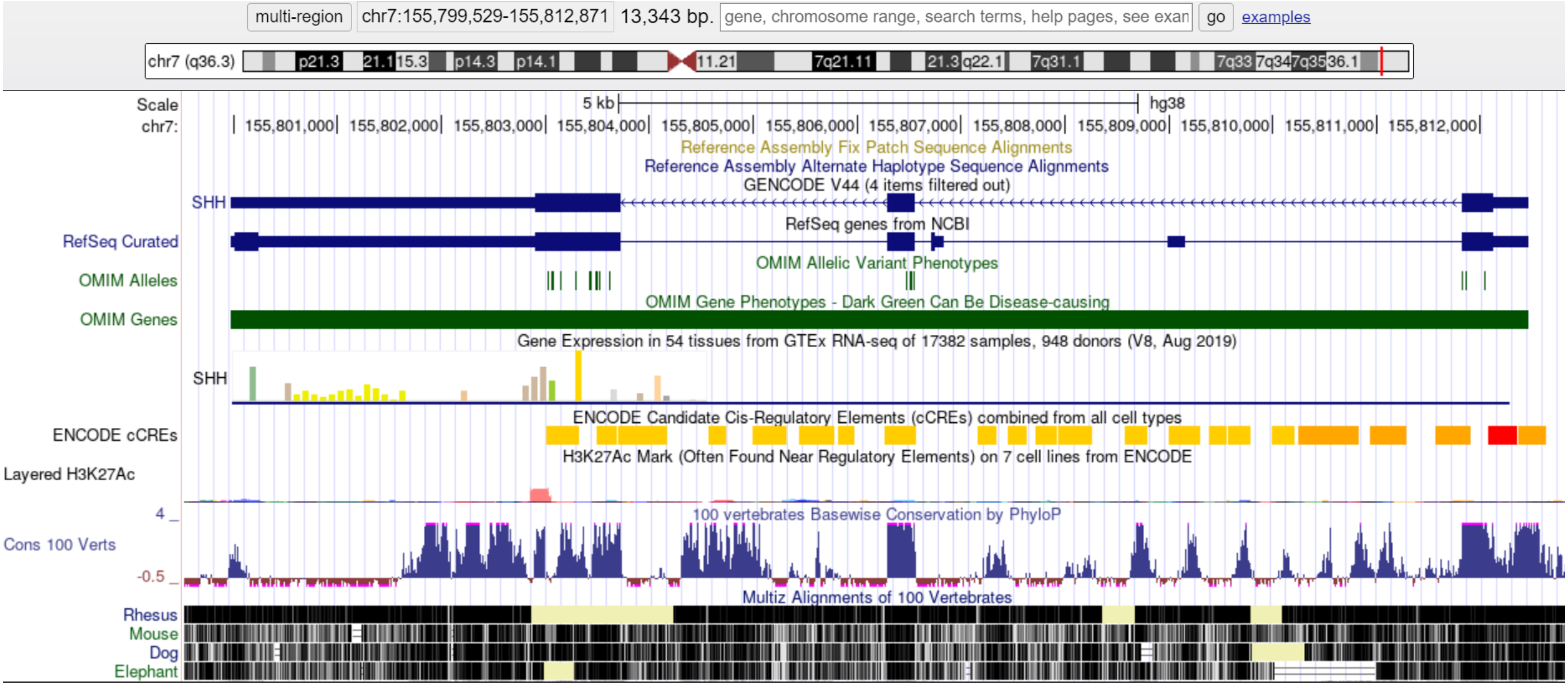
- Genome browsers are visual interfaces to genomic data
- They exist in many flavours, for different species
- Their scope is to bring together different datasets in the same framework: the reference genome

Quite some differences depending on organism:

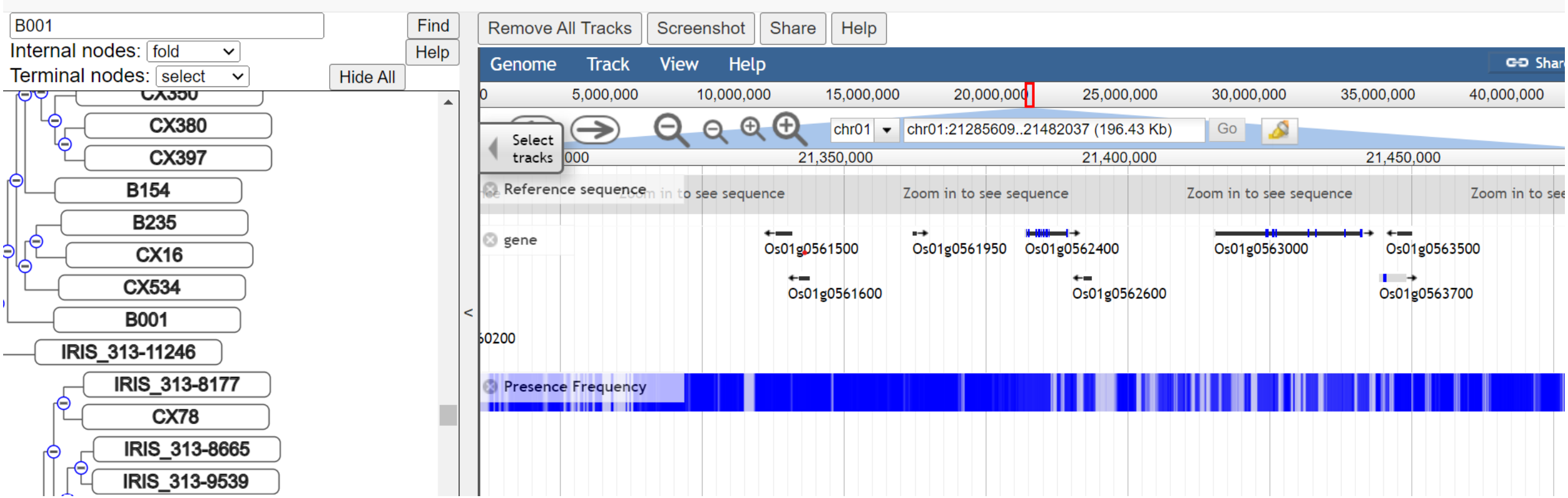
- https://genome-euro.ucsc.edu/ (human)
- https://www.gramene.org/ (plants)



Genome browsers are information dense



Browsers in a pangenomic dimension



https://cgm.sjtu.edu.cn/3kricedb/visualization/?tracks=DNA%2Cgene%2CPF&loc=chr01%3A21285609..21482037&highlight=

We now know where diversity comes from (ref. Mutation)

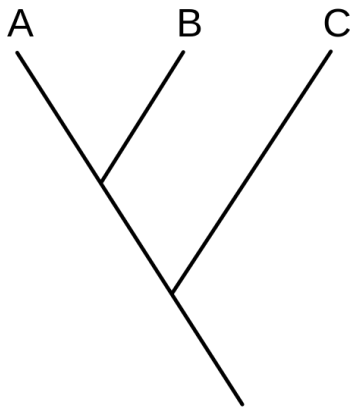
The quest of the geneticist is to **assess** it and **quantify** it

Why? It is by undestanding diversity that one may use genetic tools to e.g. improve crop traits, develop efficient and personalized cures, etc

Markers

Markers are features of organisms that can be used to assess their diversity, relatedness, and uniqueness

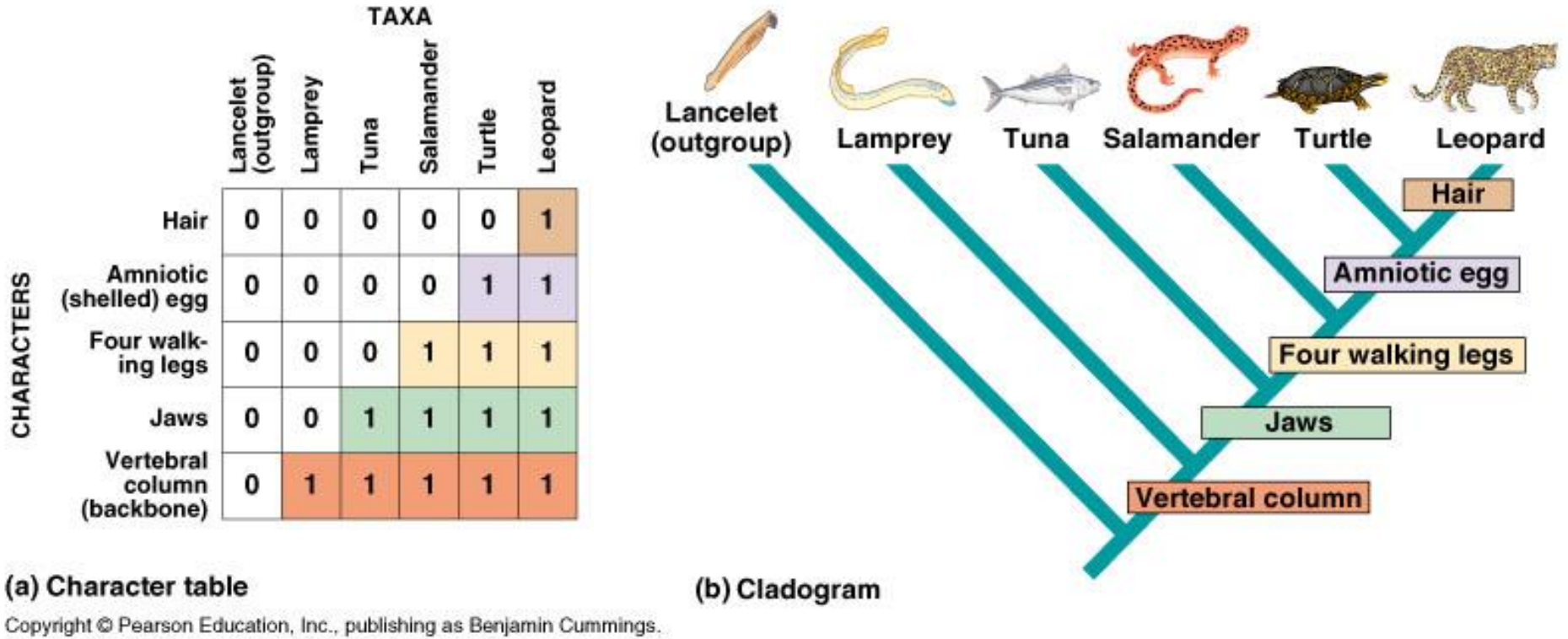
Diversity



Relatedness

Uniqueness

Multiple markers can be combined to gain information

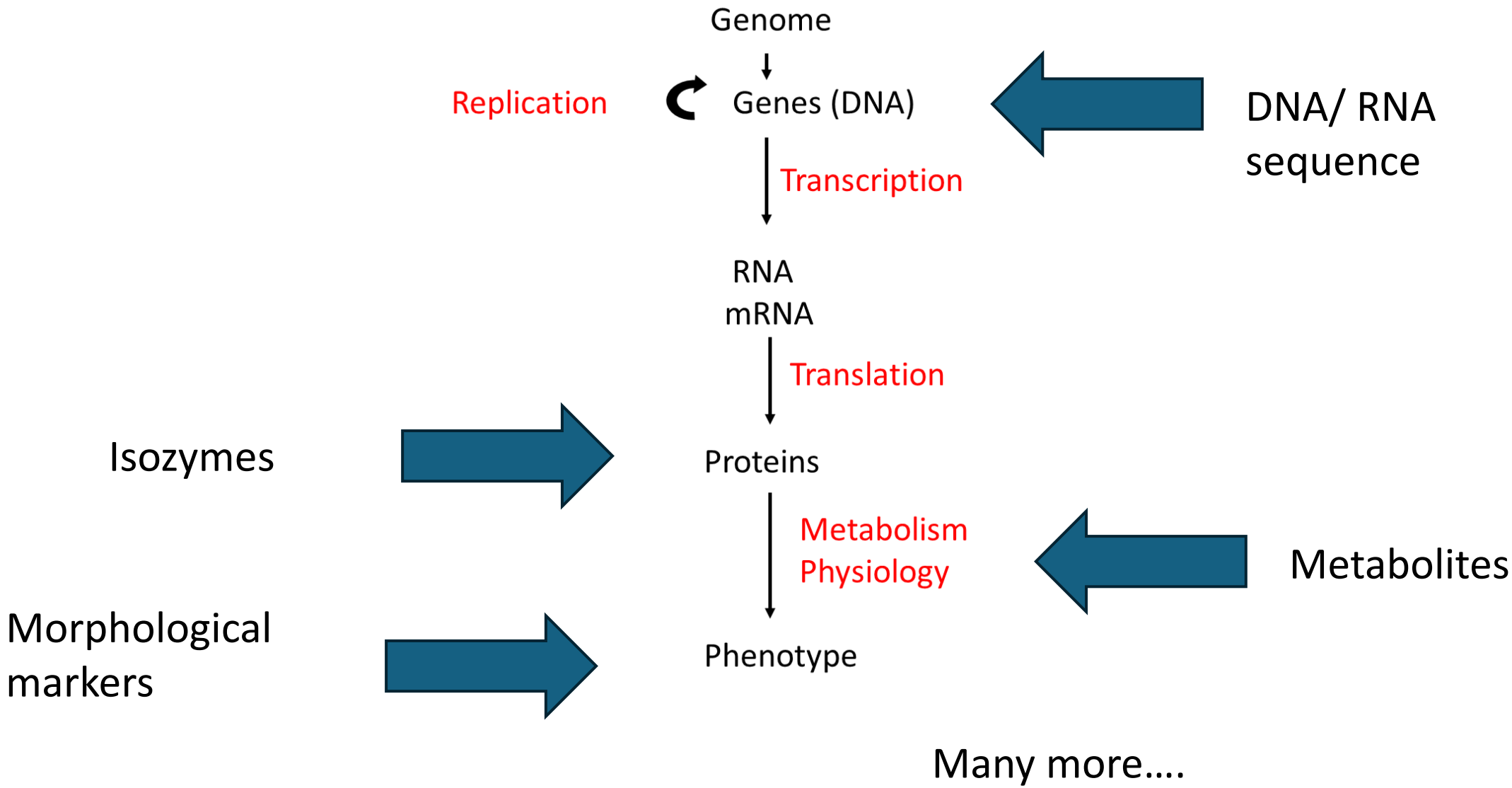


(a) Character table

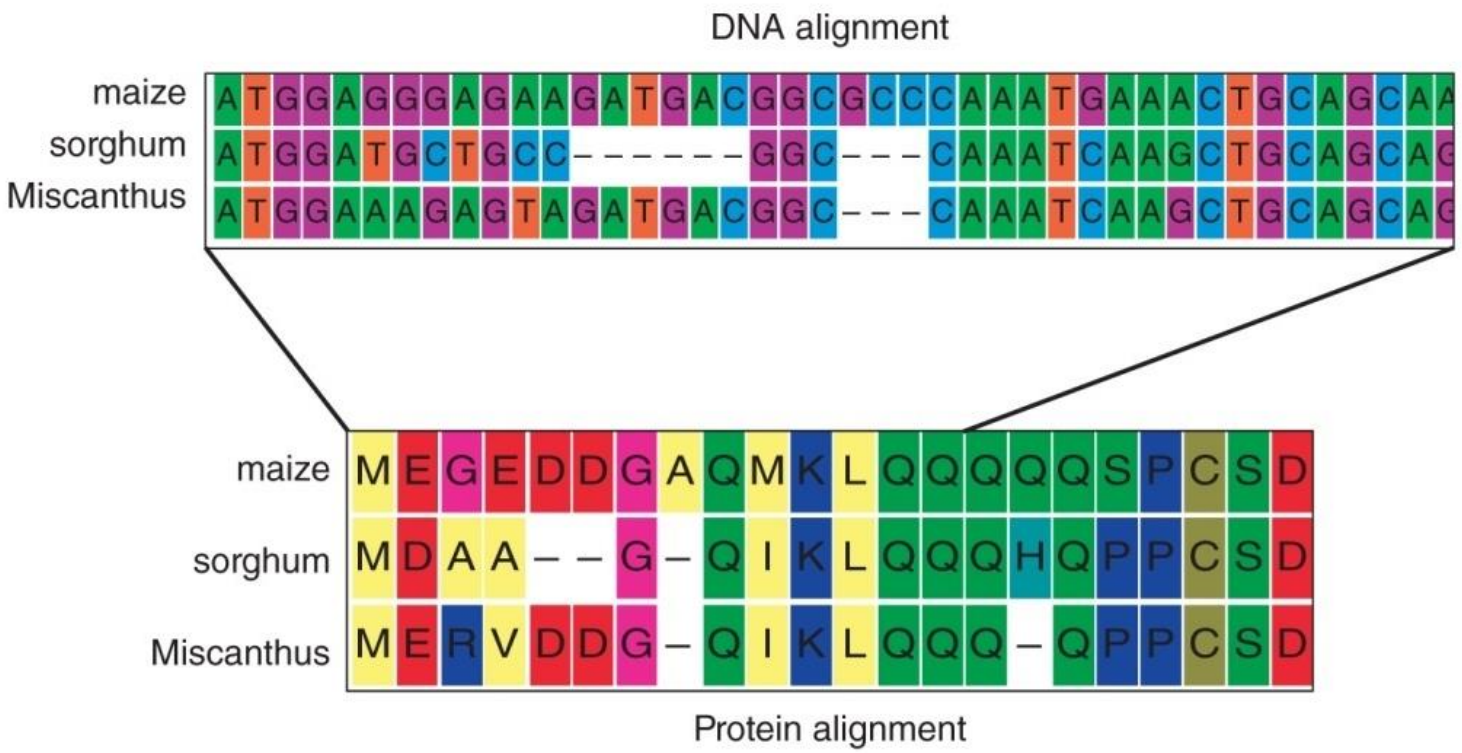
(b) Cladogram

Rules for a desirable marker:

1. Polymorphic, *i.e.* vary among individuals
2. Robust, *i.e.* should provide consistent results
3. Following a mendelian behaviour, *i.e.* follow laws of inheritance with no influence from environment
4. Informative, *i.e.* with dense information content
5. Easy to analyze, *i.e.* information should be easy to grasp
6. Inexpensive, *i.e.* can be produced cheaply in high numbers
7. Amenable to automation


Genome
Replication
Genes (DNA)
DNA/ RNA sequence
Transcription
RNA
mRNA
Translation
Isozymes
Proteins
Metabolism
Physiology
Metabolites
Morphological markers
Phenotype
Many more….

Not suprisingly, the ultimate molecular marker is a DNA marker



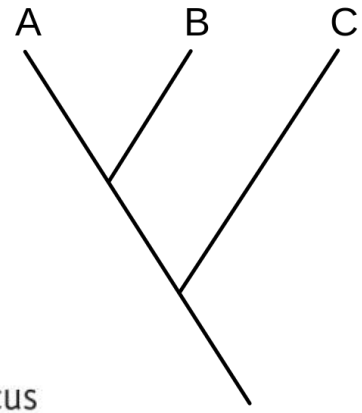
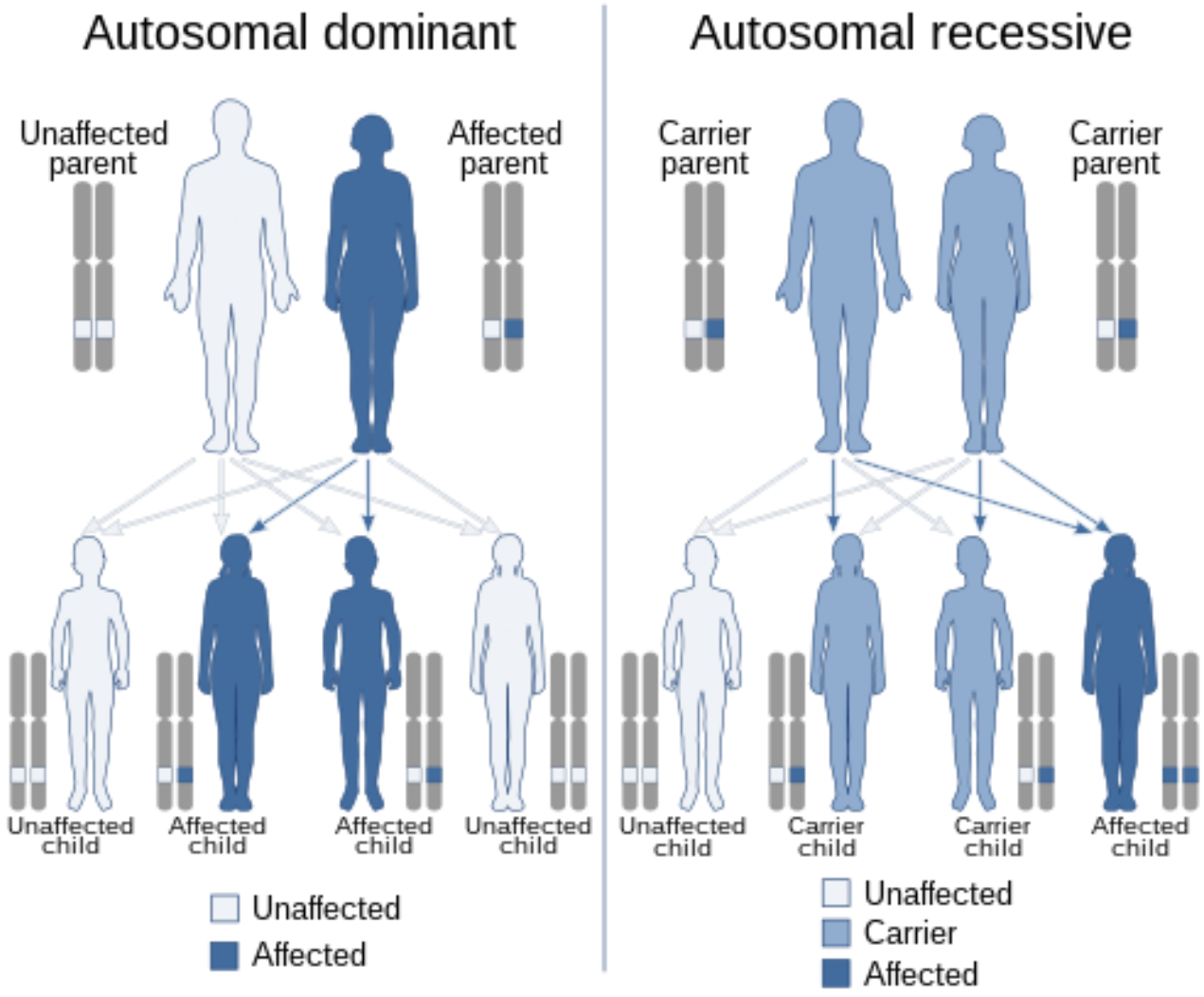


Figure 2.2 Alignment of nucleotides and the corresponding amino acids from the beginning portion of the *ramosa1* locus in maize, sorghum, and the ornamental plant *Miscanthus sinensis*. The 42 base pairs of the DNA alignment correspond to the first 14 amino acids in the protein. Two indels are visible in these first 42 base pairs, plus 9 single nucleotide polymorphisms

Good markers are polymorphic markers;

- They can identify **alleles**, i.e. alternative form of genes/loci
- Alleles can be two (biallelic markers) or more (multiallelic markers). If you have just one allele, then you don't have polymorphism
- Tipically, alleles of a gene A are denoted A, a, in classic mendelian term dominant and recessive
- Markers can be dominant or codominant; depending on whether you can detect the recessive allele or not



Morphological markers are based on the appearance (phenotype) of individuals

PLANT DATA

4.1 VEGETATIVE

4.1.1 Growth class (seasonality)

1 Winter
2 Facultative (intermediate)
3 Spring

4.1.2 Plant height

Height of plant at maturity, measured in cm from ground to top of spike, excluding awns

4.2 INFLORESCENCE

4.2.1 Days to flower

Counted as days from sowing to 50% of plants in flower. However, when planting in dry soils in dryland areas it is counted from the first day of rainfall or irrigation which is sufficient for germination

4.2.2 Spike density (see Fig. 1)

A visual measure of the density of a spike measured on a 1-9 scale. (NB. Spike density is not the same as spike shape.)

1 Very lax
3 Lax
5 Intermedite
7 Dense
9 Very dense

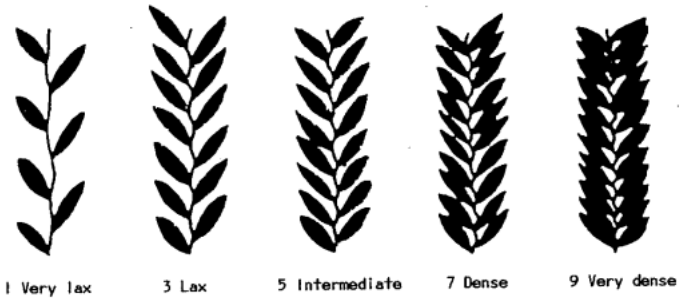


Fig. 1. Spike density

4.2.3 Awnedness

0 Awnless
3 Awnletted (short awns)
7 Awned (conspicuous awns)

4.2.4 Glume colour

Observed on the outer glume

1 White
2 Red to brown
3 Purple to black

4.2.5 Glume hairiness

Measured on outer side of sterile glume

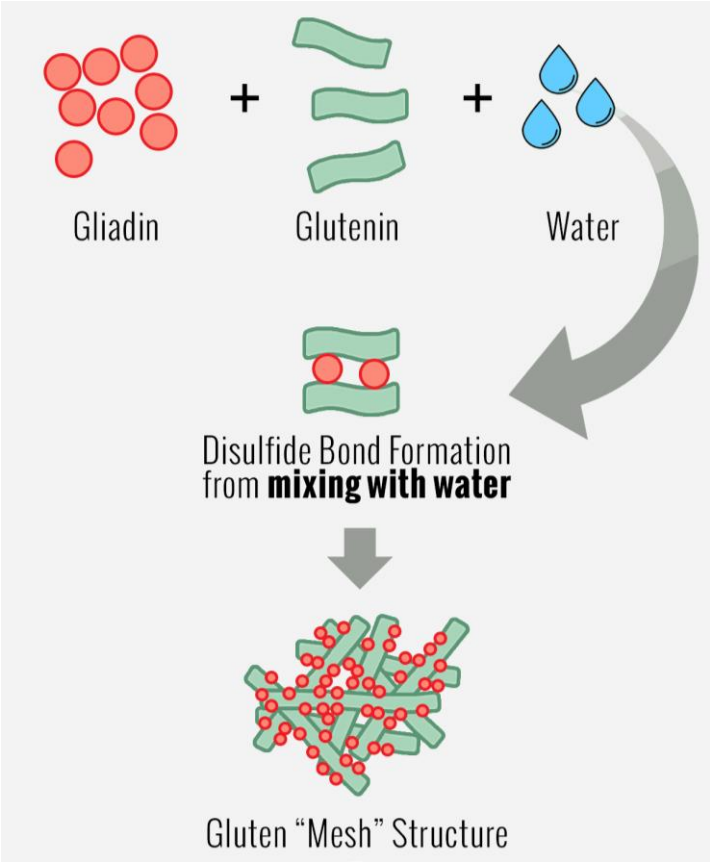
0 Absent
3 Low
7 High

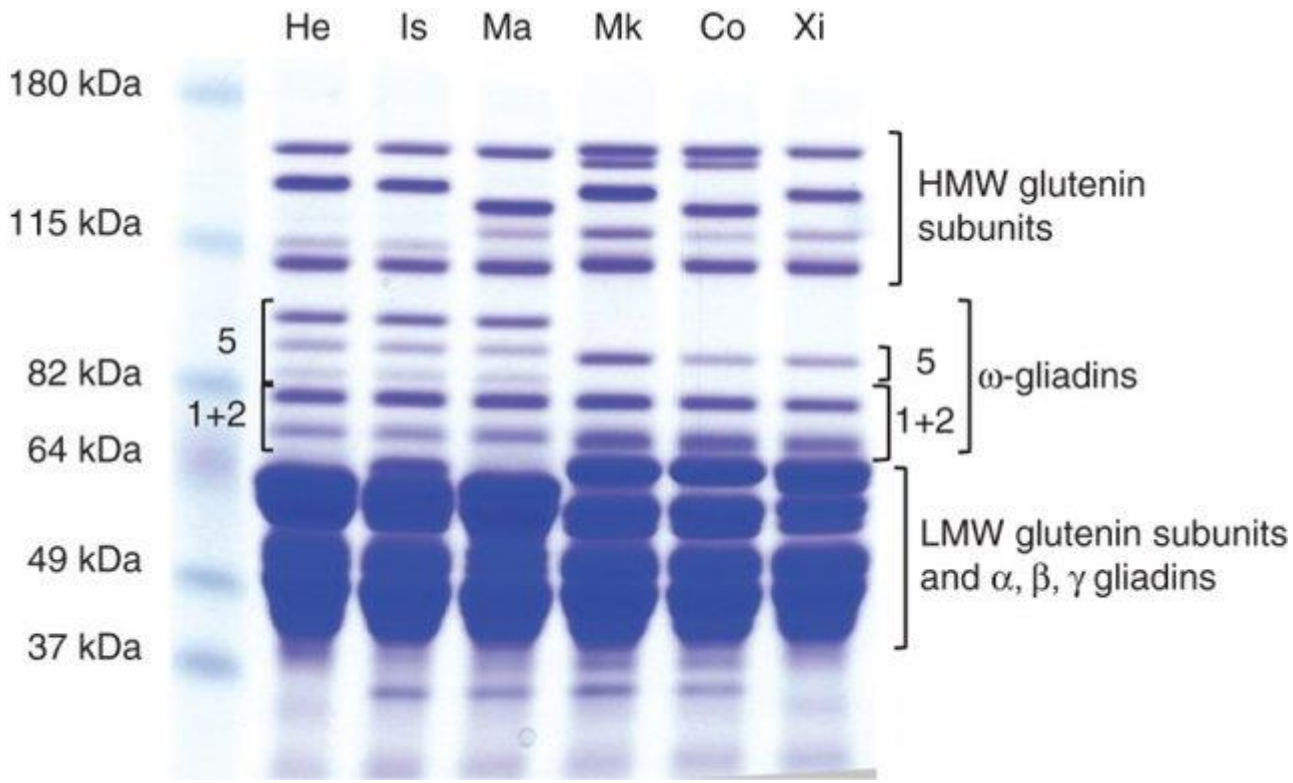
From a guide to score wheat traits for breeding

Easy to score, but far from perfect; not robust; not easy to automate

Biochemical / molecular markers

E.g. protein content of wheat gluten; different proteins / different forms of the same protein (isozymes)

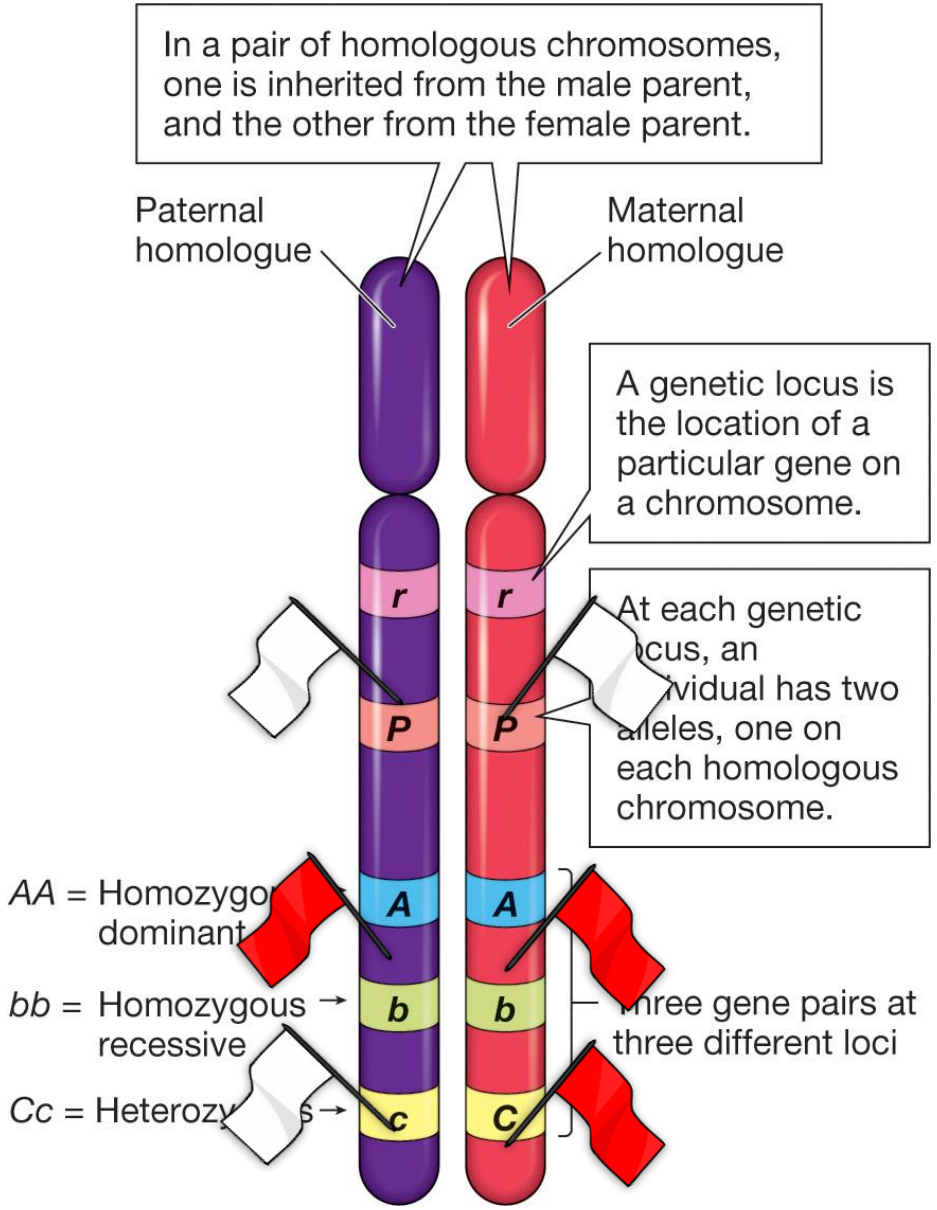




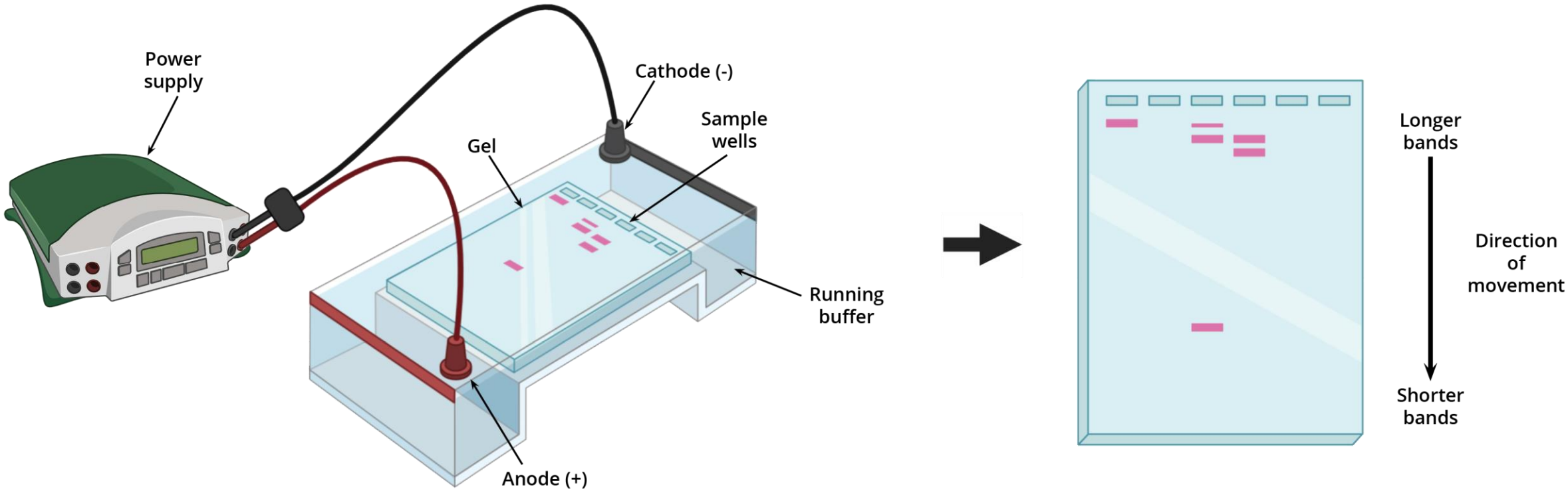
A bit more robust than morphological markers; hard to score; expensive; not informative

Genotypic markers

- Markers that can be used to characterize variation across the genome, and to tell apart individual alleles
- Each identifying unambigously a specific chromosomal region (a locus)
- Markers are tools, not the objective; an indirect approach to assess diversity at the DNA level

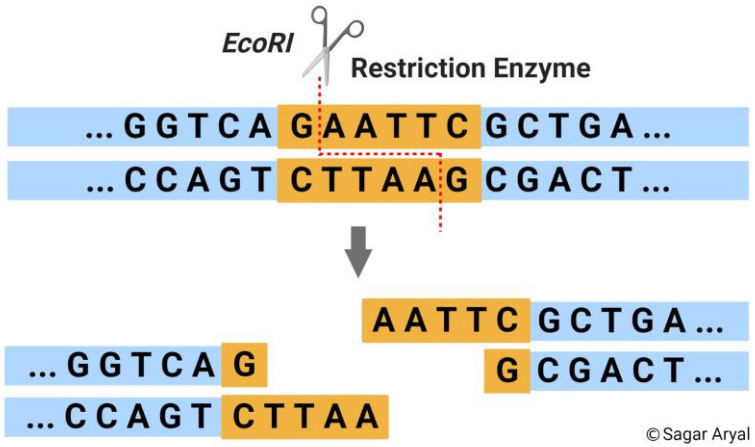


Key technology: electrophoresis



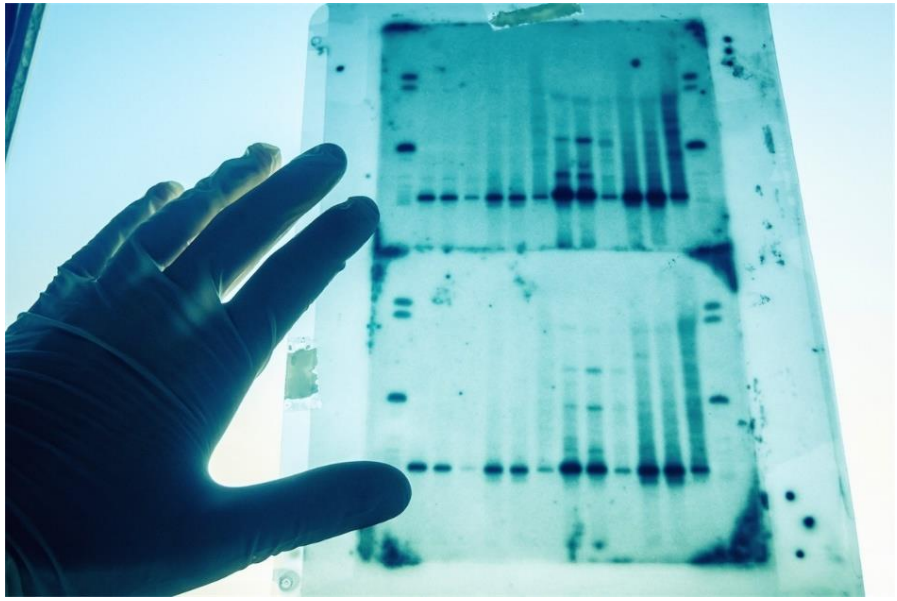
Key technology: restriction enzymes

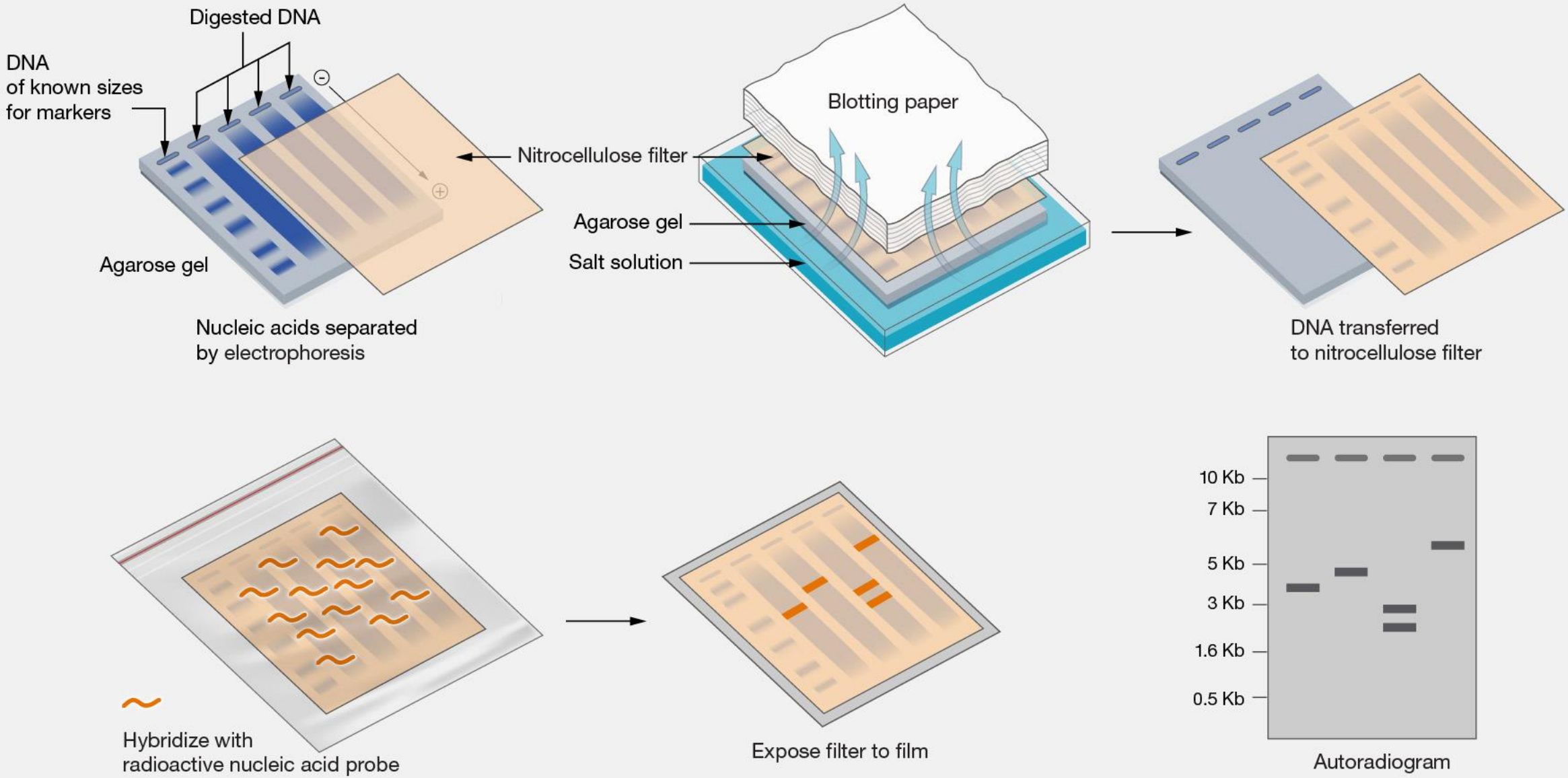
- A protein isolated from bacteria that cleaves DNA sequences at sequence-specific sites, producing DNA fragments with a known sequence at each end
- There are at least three thousand of them
- Each one of these enzymes cuts a specific DNA sequence and doesn't discriminate as to where the DNA comes from



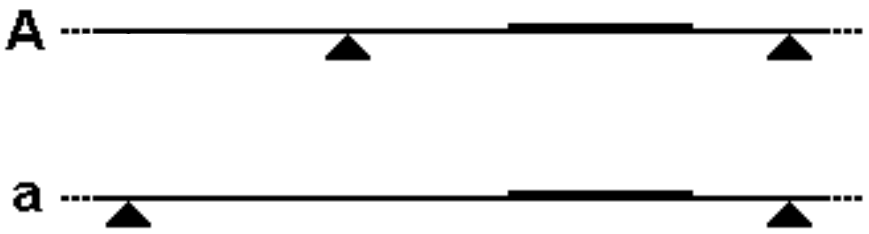
Restriction Fragment Length Polymorphism (RFLP)

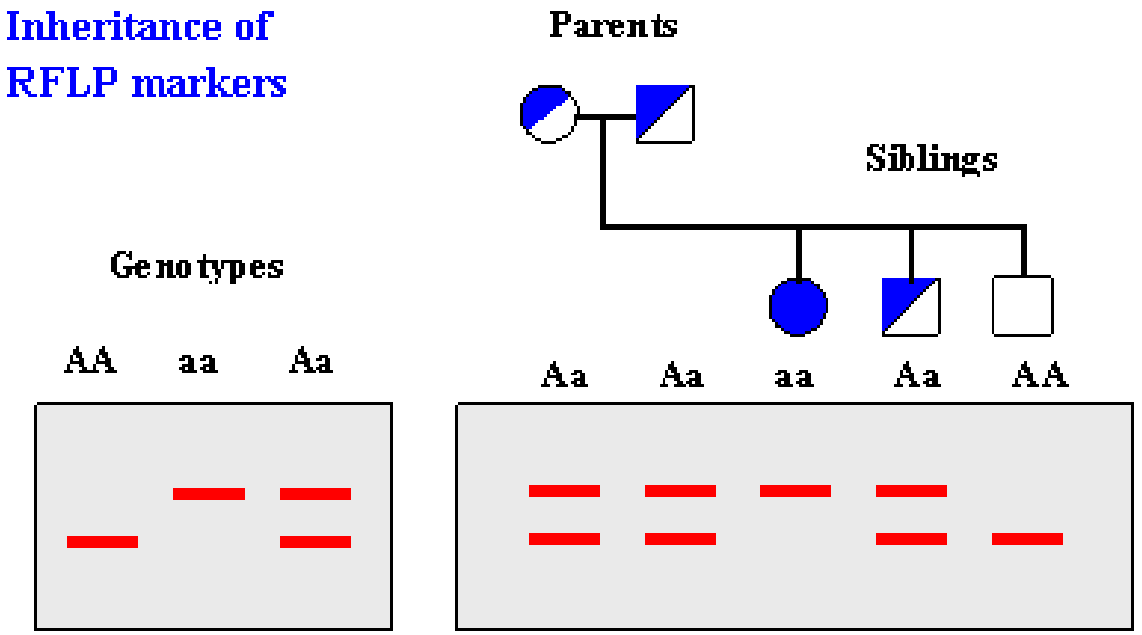
- A sample of DNA is fragmented with the application of a restriction enzyme
- DNA fragments are then separated by length through gel electrophoresis and transferred to a membrane via Southern blot
- Hybridization of the membrane to a labeled DNA probe, representing the locus of interest, determines the length of the fragments which are complementary to the probe




Digested DNA
DNA
of known sizes
for markers
Agarose gel
Nucleic acids separated
by electrophoresis
Nitrocellulose filter
Blotting paper
Agarose gel
Salt solution
DNA transferred
to nitrocellulose filter
Hybridize with
radioactive nucleic acid probe
Expose filter to film
10 Kb
7 Kb
5 Kb
3 Kb
1.6 Kb
0.5 Kb
Autoradiogram

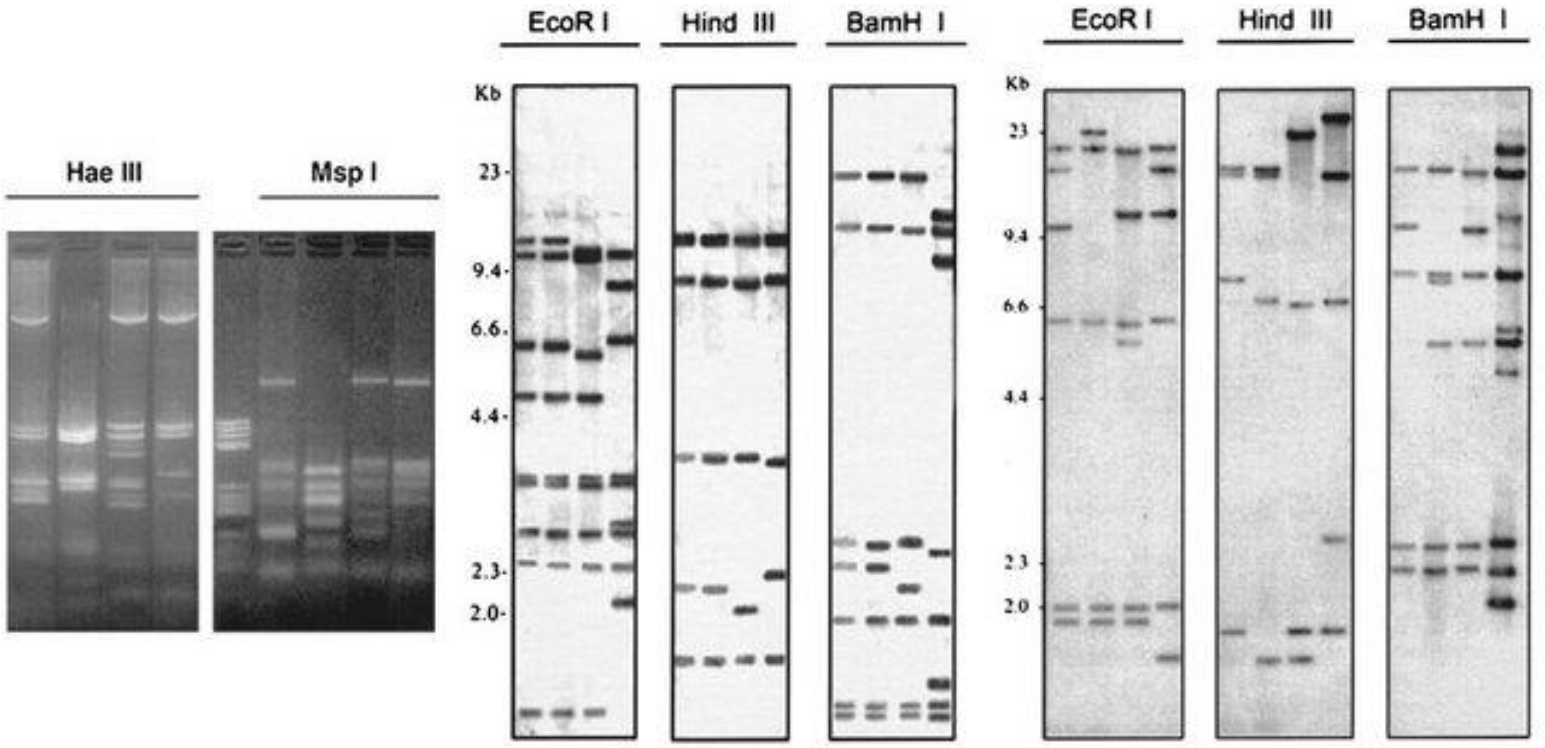
- An RFLP occurs when the length of a detected fragment varies between individuals, indicating non-identical sequence homologies (based on RE cut site)
- Each fragment length is considered an allele, whether it actually contains a coding region or not, and can be used in subsequent genetic analysis.





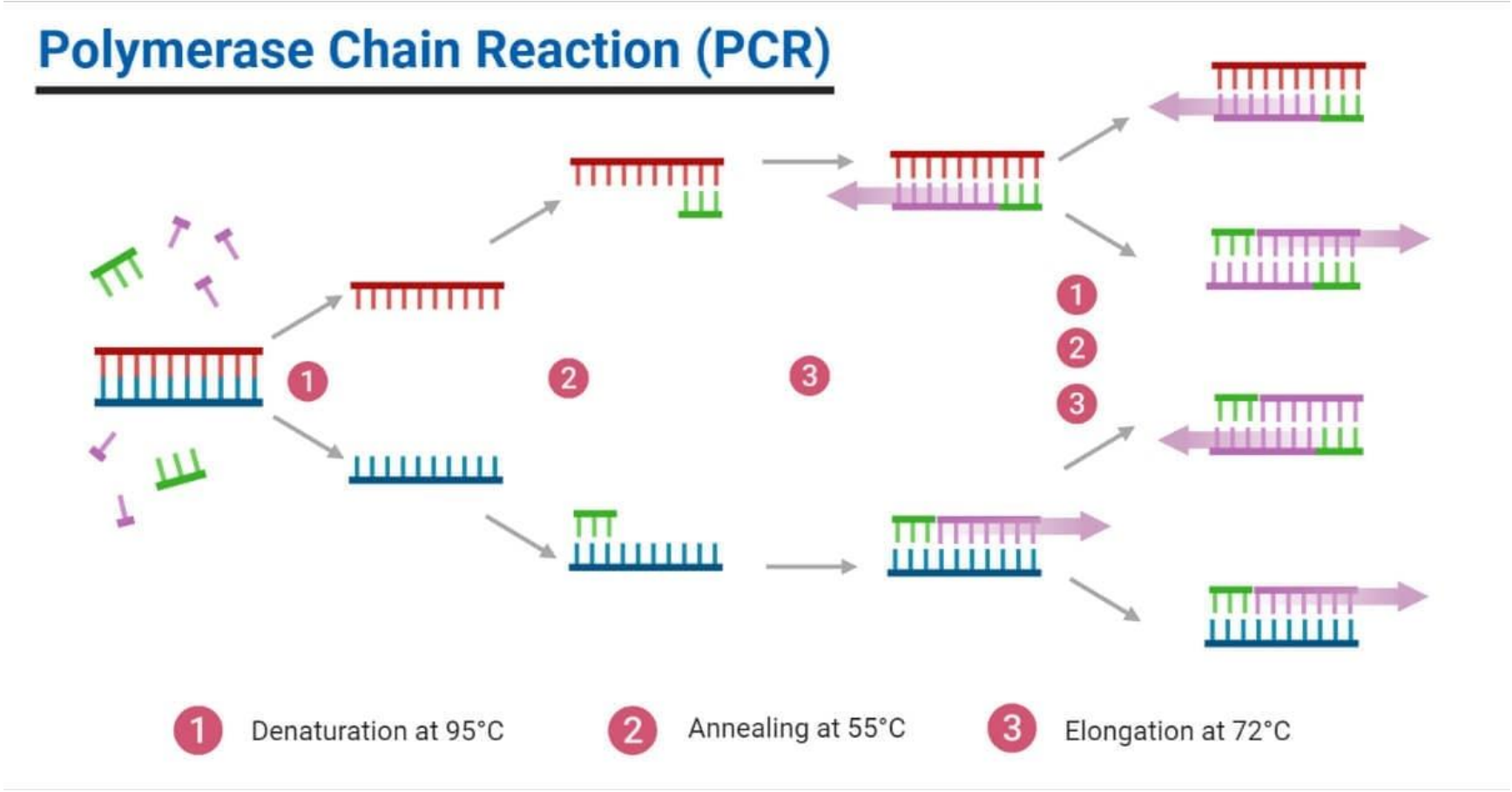
Good: RFLP markers are very specific and repeatable; they are co-dominant; you can combine different RE to get a better picture (each locus may have more than one RE site)

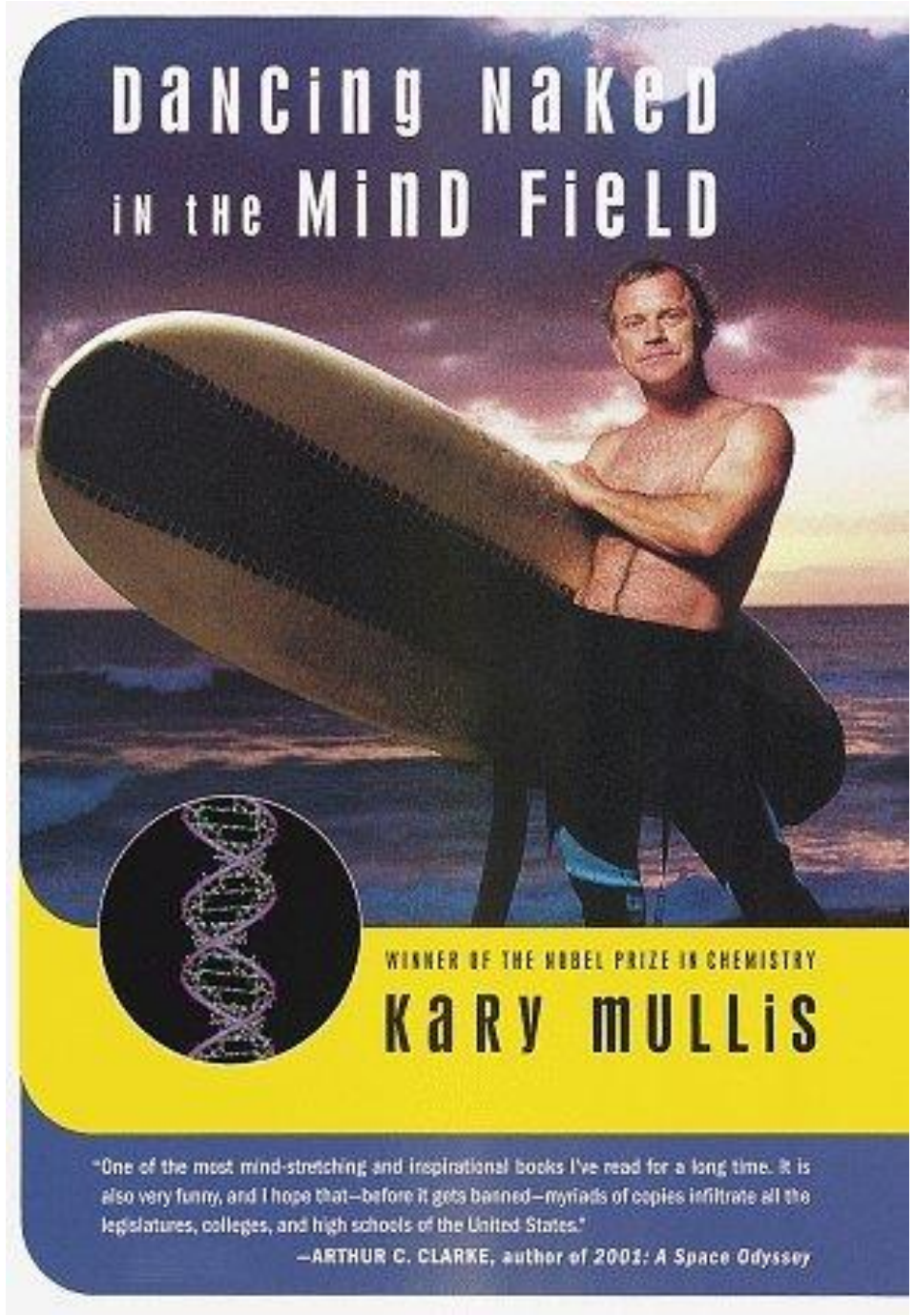
Bad: the technique is slow and cumbersome. It requires a large amount of sample DNA, and the probe labeling, DNA fragmentation, electrophoresis, blotting, hybridization, washing, and autoradiography can take up to a month to complete.



- Type: co-dominant, locus specific (due to throughput limitations)
- Number of markers: dozens
- Number of alleles: 2

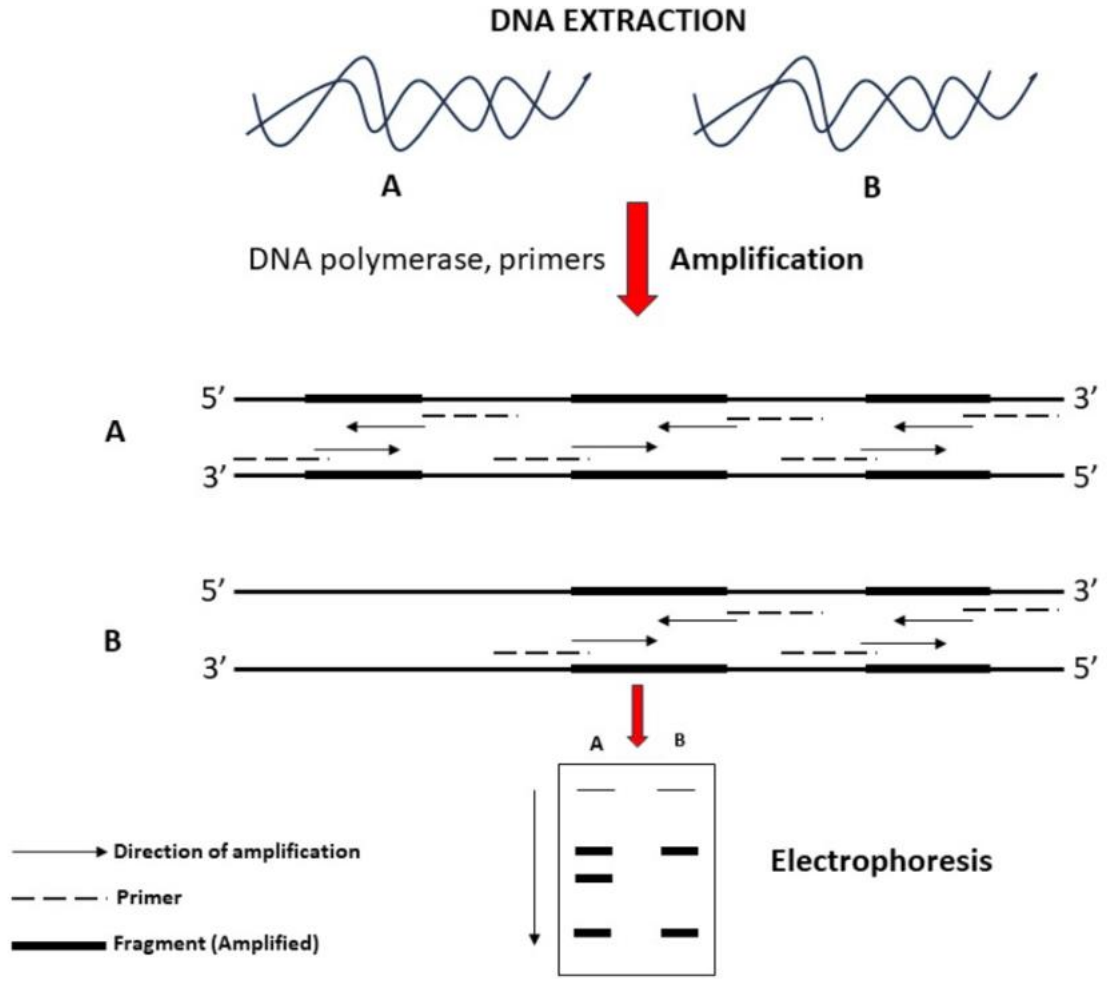
Key technology: PCR




Dancing Naked
in the Mind Field
WINNER OF THE NOBEL PRIZE IN CHEMISTRY
Kary Mullis
"One of the most mind-stretching and inspirational books I've read for a long time. It is also very funny, and I hope that—before it gets banned—myriads of copies infiltrate all the legislatures, colleges, and high schools of the United States."
—ARTHUR C. CLARKE, author of 2001: A Space Odyssey

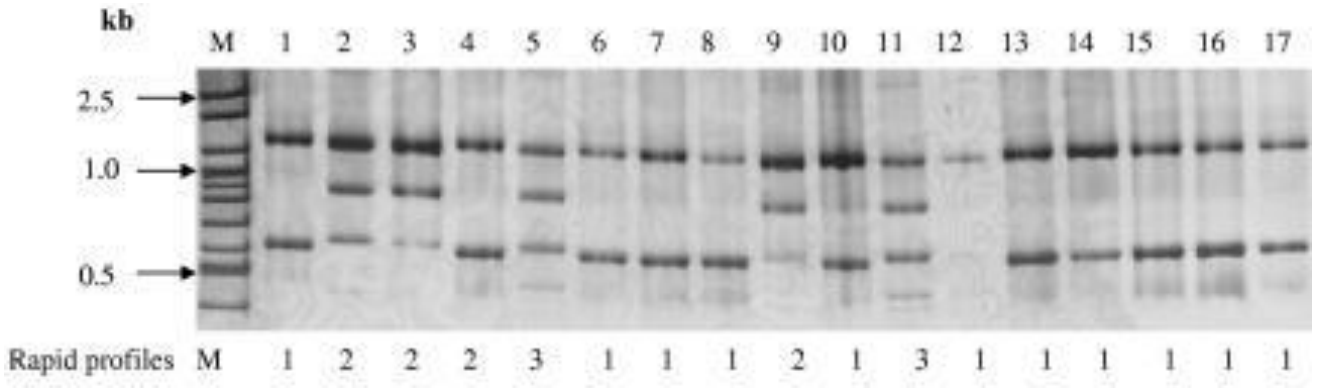
Random Amplified Polymorphic DNA (RAPD)

- Based on PCR amplification (not restriction)
- Identical 10-mer primers will or will not amplify a segment of DNA, depending on positions that are complementary to the primers' sequence and distance of complementary sequences
- The presence/absence of PCR products reveals the polymorphism



Bad: dominant; based on PCR reaction and lab-dependent; mismatch in primer lower efficiency

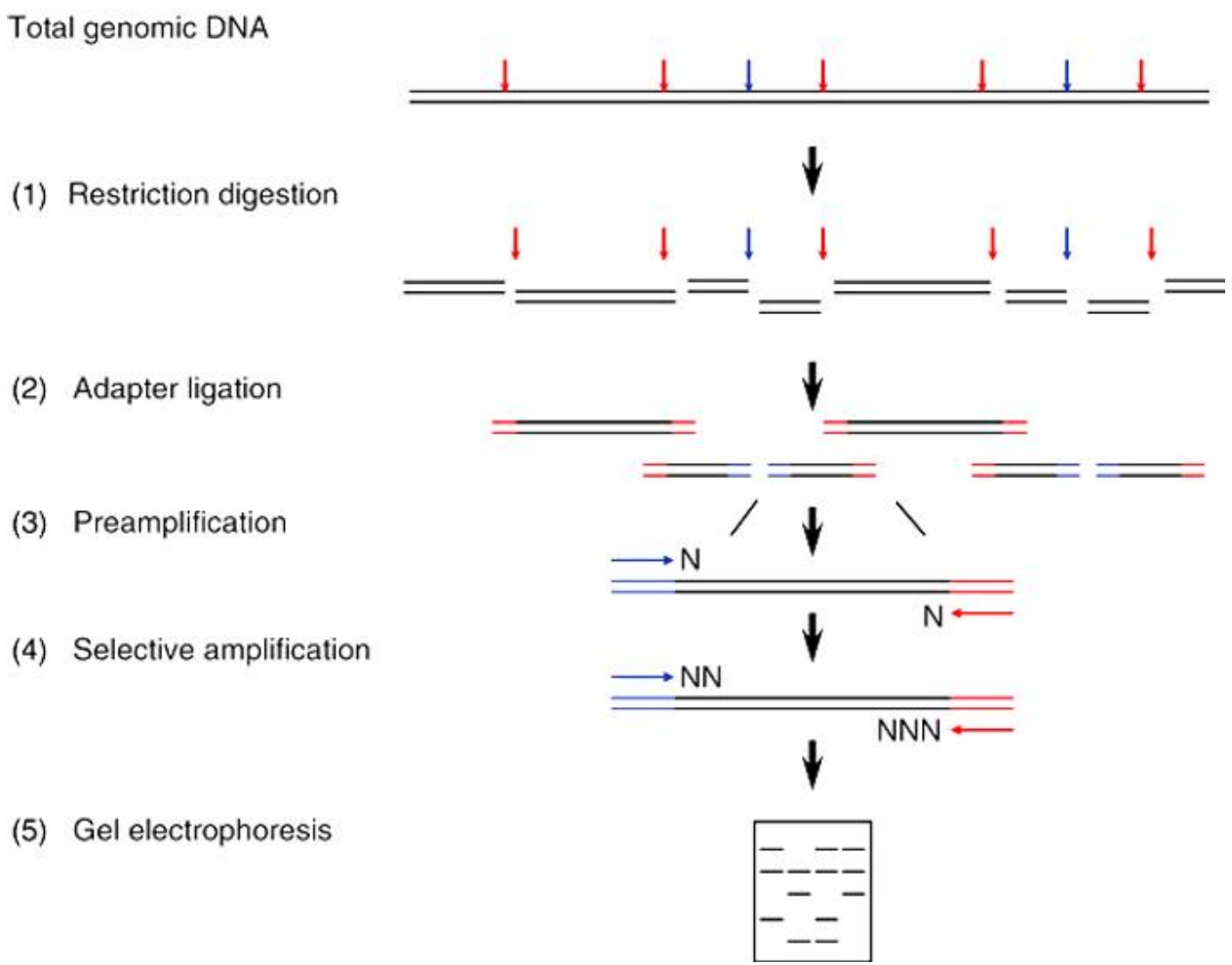
Good: cheap, quick, polymorphic bands can be isolated and sequenced to derive specific markers (SCAR)



- Type: dominant, genome-wide
- Number of markers: tens to hundreds
- Number of alleles: 2

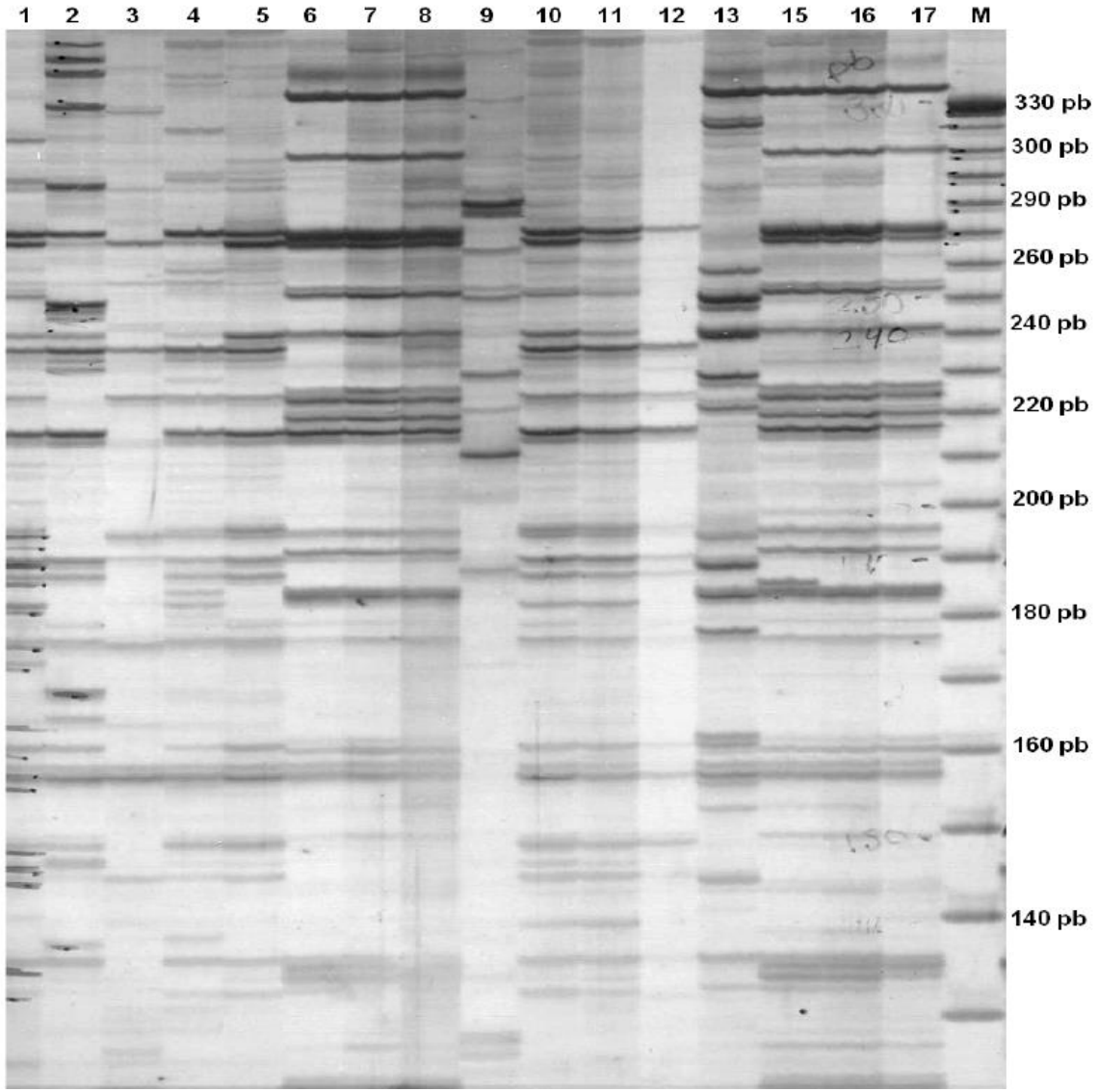
Amplified Fragment Length Polymorphism (AFLP)

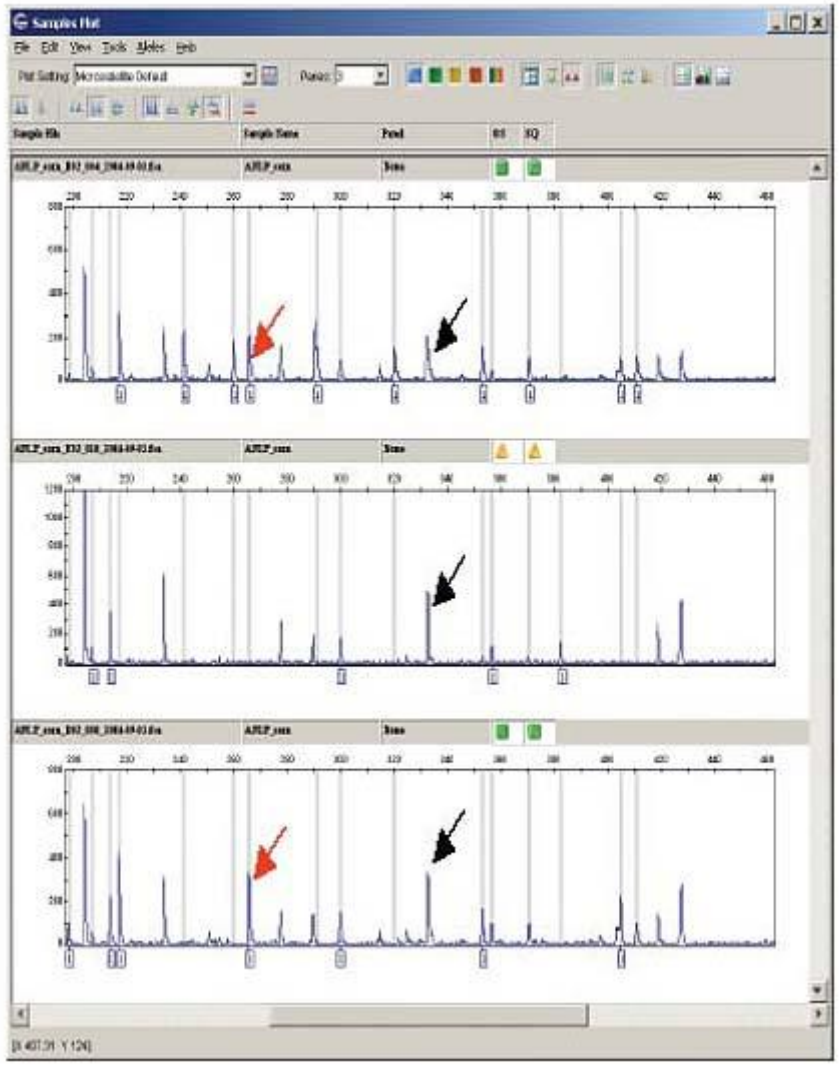
Very similar to RFLP but going through an amplification phase (less DNA input, more markers per experiment)



No prior information on the sequence is required!!!

- Type: dominant, genome-wide
- Number of markers: hundreds
- Number of alleles: 2 (presence/absence)

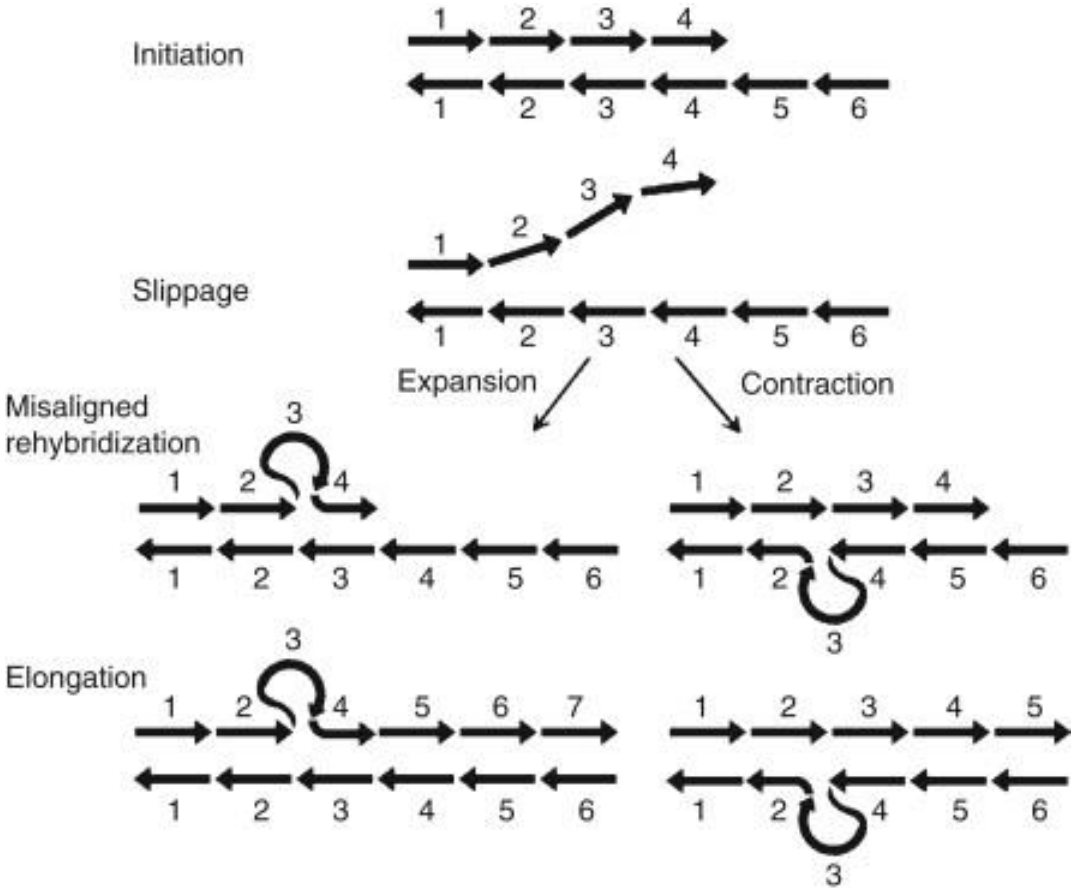




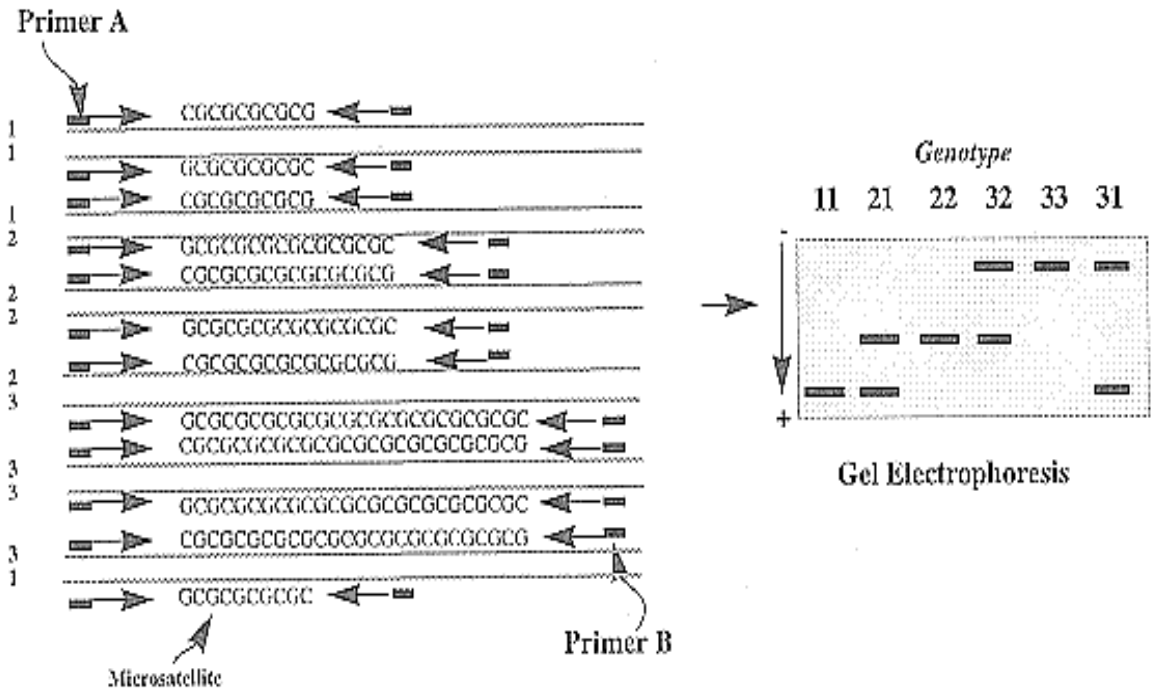
Band (polymorphism) calling via gel or via capillary separation

Single sequence repeats (SSR) or microsatellites

- SSR or microsatellites are loci with repeats of a motif sequence 1-6 bp in length, ubiquitous in the genome
- Due to this structure SSRs frequently undergo mutations, mainly due to DNA polymerase errors, which involve the addition or subtraction of a repeat unit



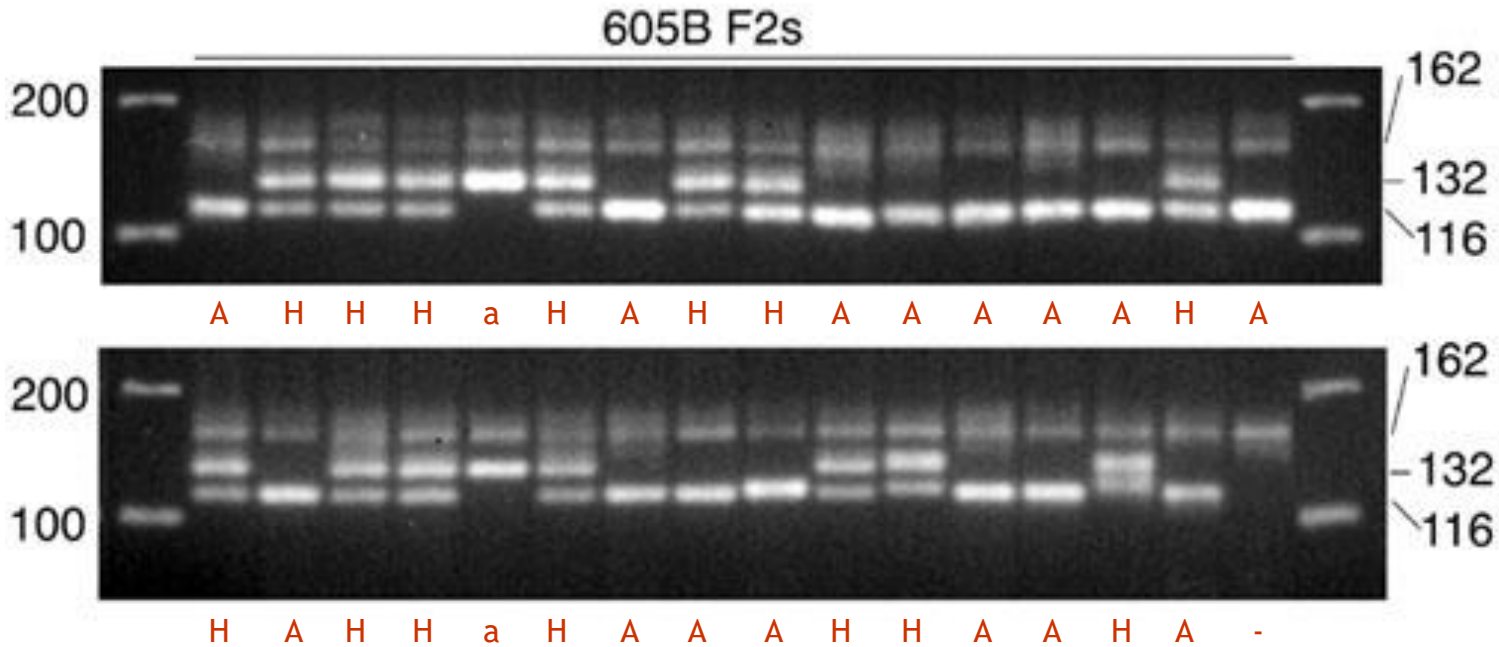
- Not every part of the genome varies with the same frequency in the population
- Non-coding DNA regions are (not) constrained by selection, and hence display higher variation in the population
- Primers can be designed on conserved regions and used to amplify SSRs; polymorphisms are given by variation in length of the PCR product



Bad: each SSR primer pair unique; not really prone to automation

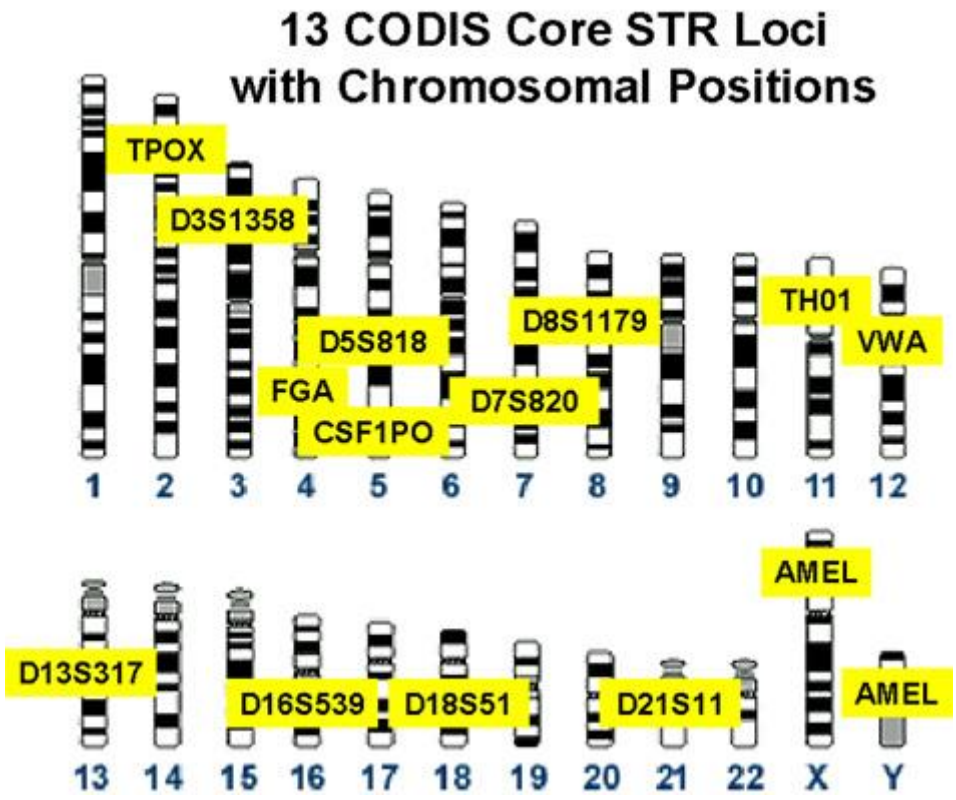
Good: cheap, highly informative, highly repeatable

- Type: co-dominant, locus specific
- Number of markers: tens to hundreds
- Number of alleles: **2+**


605B F2s
200
100
162
132
116
A H H H a H A H H A A A A A H A
200
100
162
132
116
H A H H a H A A A H H A A H A -

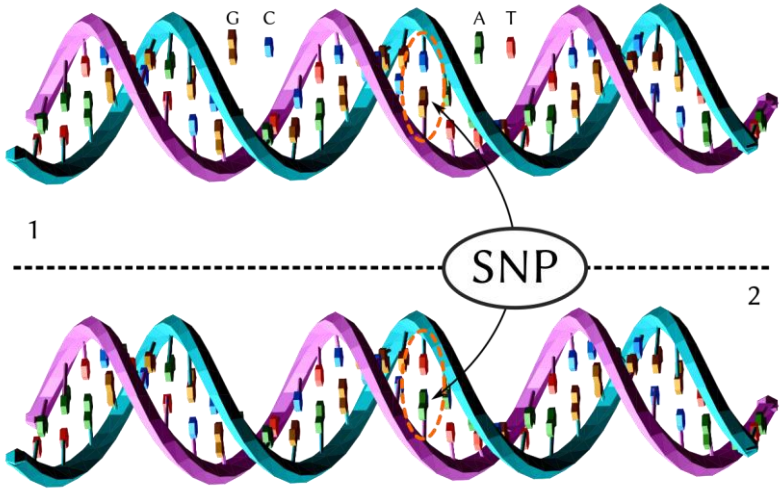
The current methodology for DNA profiling is called CODIS (Combined DNA Index System) and based on SSR

- It makes use of 20 microsatellites with sufficient variability to give only a one in 10^{18} chance that two individuals, other than identical twins, have the same profile
- As the world population is around 8×10^9, the statistical likelihood of two individuals on the planet sharing the same profile is so low as to be considered implausible when DNA evidence is presented in a court of law



Single Nucleotide Polymorphisms (SNPs)

- The ultimate source of variation: differences in DNA sequence
- Polymorphisms derive from a germline substitution of a single nucleotide at a specific position in the genome that is present in a sufficiently large fraction of considered population
- SNPs occur in non-coding regions more frequently than in coding regions. Other factors, like genetic recombination and mutation rate, can also determine SNP density
- Became possible only once DNA sequencing became reality



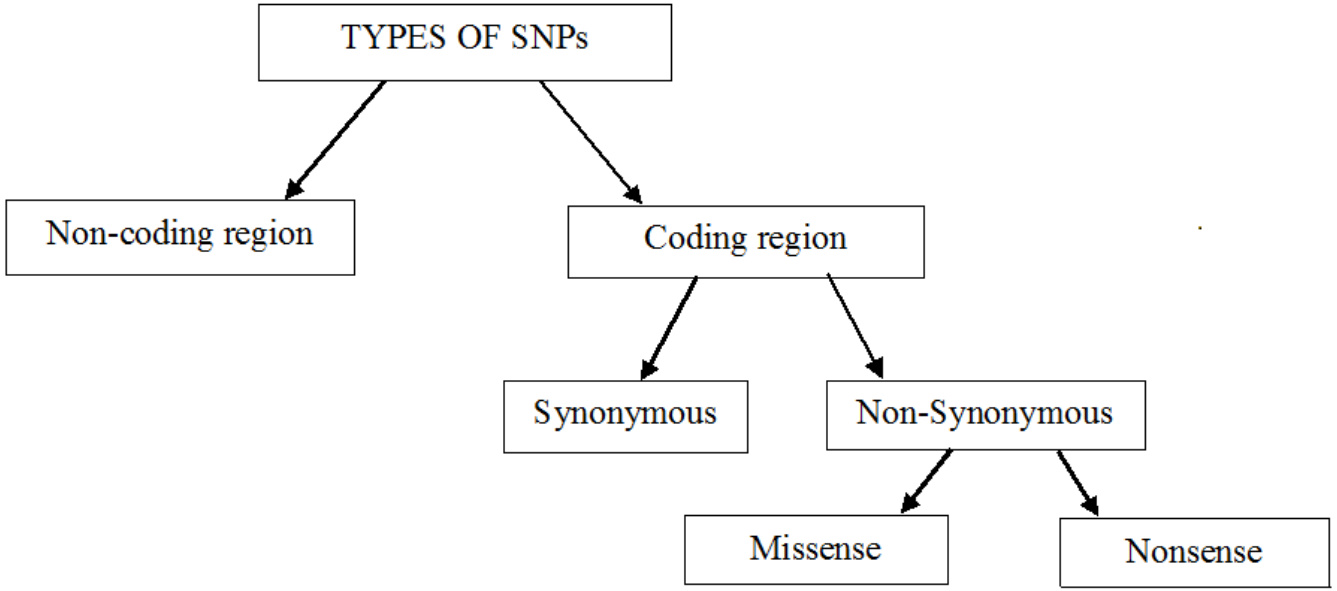
To be identified, SNPs require sequence alignment

Good: very amenable to automation, high numbers with a decreasing cost; ultimate source of information

Bad: the information per SNP is limited; data size may be daunting

- Type: co-dominant, genome wide
- Number of markers: thousands to millions
- Number of alleles: 2

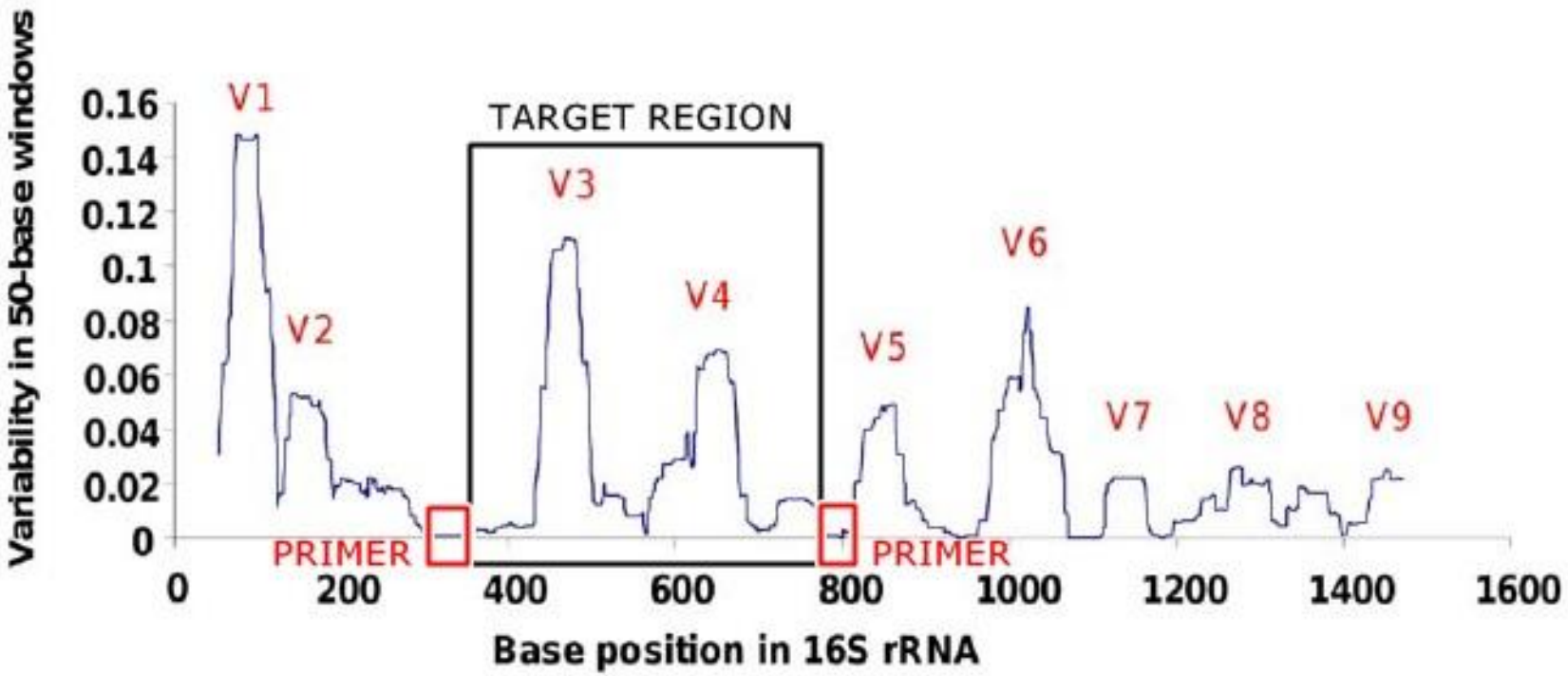
- Most SNPs are genetically neutral
- However SNPs can reflect distinguishing characteristics (0.1% affect protein function)
- SNPs can serve as markers for phenotypic traits



SNP markers for phylogenetics

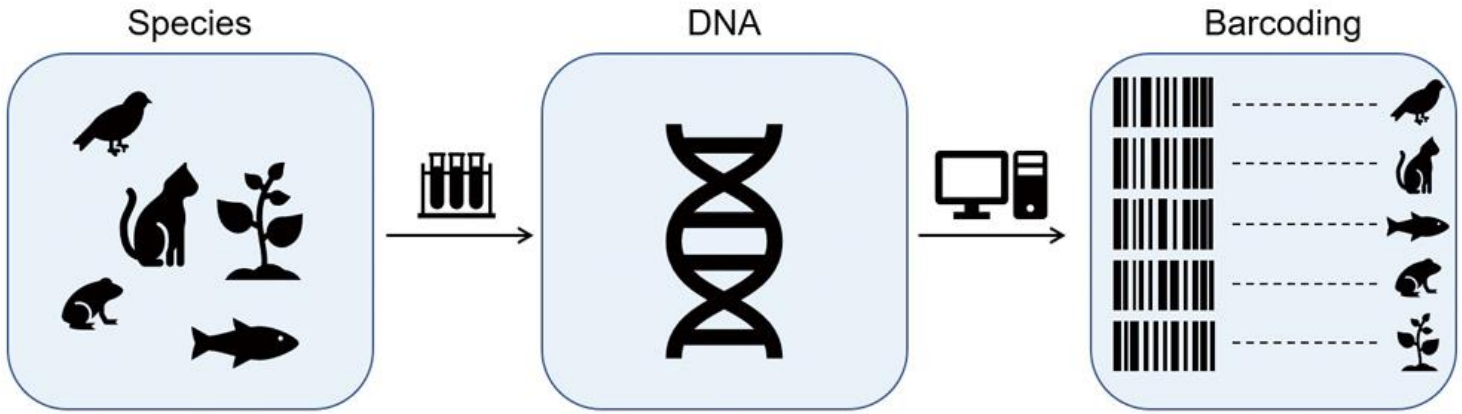
Conceptually similar to SSR, but coupled with sequencing

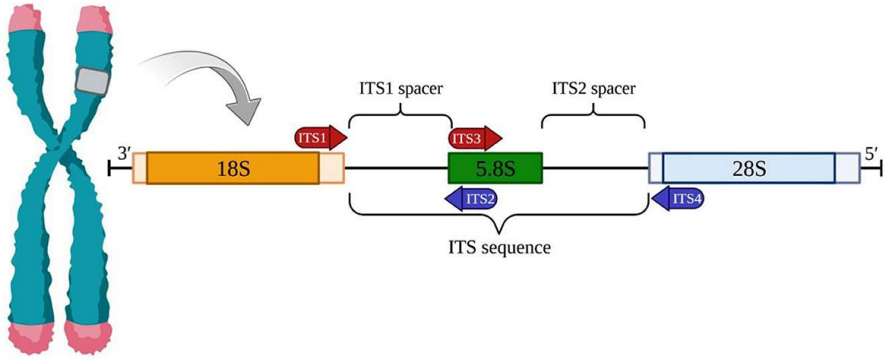
1. Design primers on invariant loci that are conserved across samples
2. Amplify (or sequence) hypervariable regions



The extent of variability of the region gives you the taxonomic relevance of the marker:

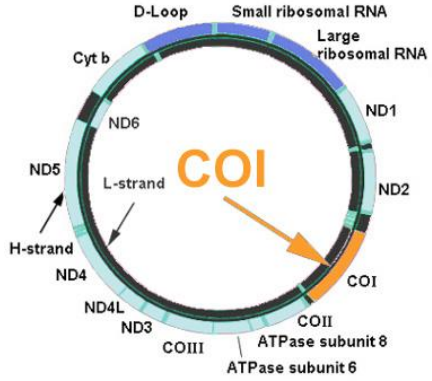
1. Low frequency of variation (=higher sequence conservation): lower taxonomic levels, *e.g.* families, phyla
2. High frequency of variation (=lower sequence conservation): higher taxonomic level, *e.g.* species, individuals





Different markers:

- ITS, COI, 16s, 18s, etc



Presence Absence Variants (PAVs) and Copy Number Variants (CNVs)

- As genomic technology advances, it is becoming increasingly clear that genomes vary also with regards to gene content and, more generally, structural variants
- As genomes become better and better characterized in the pangenomic era, PAVs and CNVs can be used as markers (and often have biological function)