Molecular methods for detecting genetic variation in organisms

KEY POINT: Genetic variation may be under selection or selectively neutral

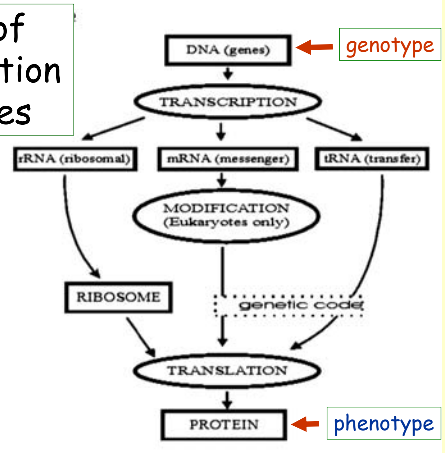
GENETIC MARKERS ARE USUALLY ASSUMED TO BE SELECTIVELY NEUTRAL

POPULATION GENETIC AND PHYLOGENETIC ANALYSES DEPEND UPON THE IDEA OF **IDB**

Flow of information in genes

Non-coding DNA

Molecular markers



Levels of organization for markers

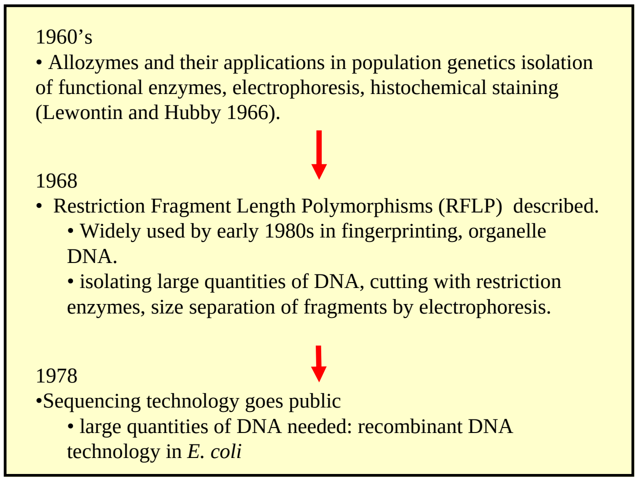
- Phenotypic variation
  - looking at what's expressed
  - structural genes, enzymes
- Examining functional genes
  - nuclear genome, organelles of Eukaryotes, (mtDNA, cpDNA), Prokaryotic plasmids
- Non-coding DNA
  - single copy (e.g. introns of Eukaryotic genes)
  - repetitive DNA's
    - (Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs), microsatelltes, intergenic spacers and intergenic spacers in rRNA genes)
  - single nucleotide polymorphisms (SNP)

The first millennium: scope of population genetic studies defined by the available technology

The Dark Ages until the late 1800's, early 1900's
- Morphological variants
  - Mendel's peas, chickens, *Paramecium*, fruit flies, maize
  - coat color in mice, guinea pigs, horses
  - shell patterns in *Cepoea nemoralis* (escagaux)
  - blood groups
  - disease susceptibility in humans & domestic animals

1930's
- Chromosomal inversion polymorphisms
  - isolating and staining chromosomes

1950's
- Biochemical polymorphisms (discovery of DNA structure)



1960’s

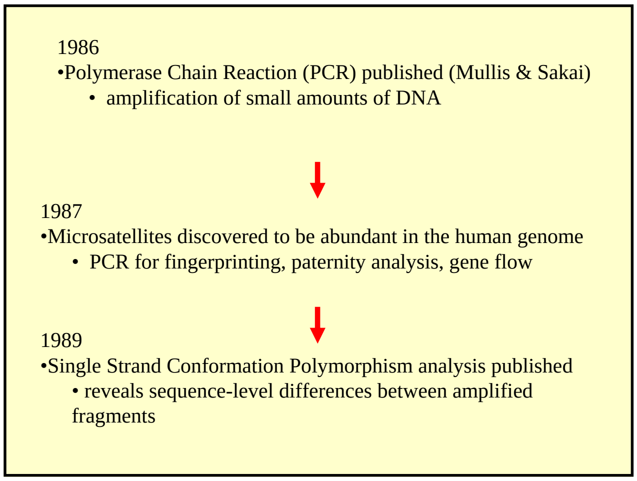
- Allozymes and their applications in population genetics isolation of functional enzymes, electrophoresis, histochemical staining (Lewontin and Hubby 1966).

1968

- Restriction Fragment Length Polymorphisms (RFLP) described.
  - Widely used by early 1980s in fingerprinting, organelle DNA.
  - isolating large quantities of DNA, cutting with restriction enzymes, size separation of fragments by electrophoresis.

1978

- Sequencing technology goes public
  - large quantities of DNA needed: recombinant DNA technology in *E. coli*



1986

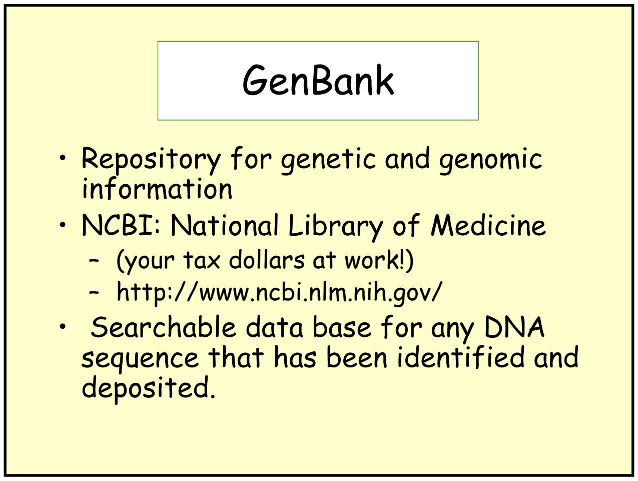
- Polymerase Chain Reaction (PCR) published (Mullis & Sakai)
  - amplification of small amounts of DNA

1987

- Microsatellites discovered to be abundant in the human genome
  - PCR for fingerprinting, paternity analysis, gene flow

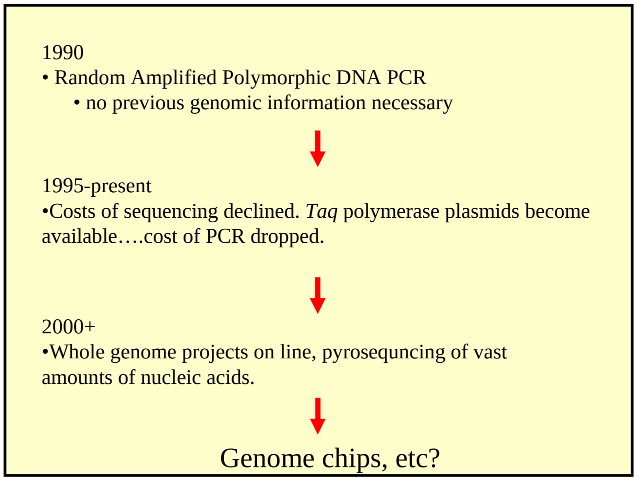
1989

- Single Strand Conformation Polymorphism analysis published
  - reveals sequence-level differences between amplified fragments



GenBank

- Repository for genetic and genomic information
- NCBI: National Library of Medicine
  - (your tax dollars at work!)
  - http://www.ncbi.nlm.nih.gov/
- Searchable data base for any DNA sequence that has been identified and deposited.



1990

- Random Amplified Polymorphic DNA PCR
  - no previous genomic information necessary

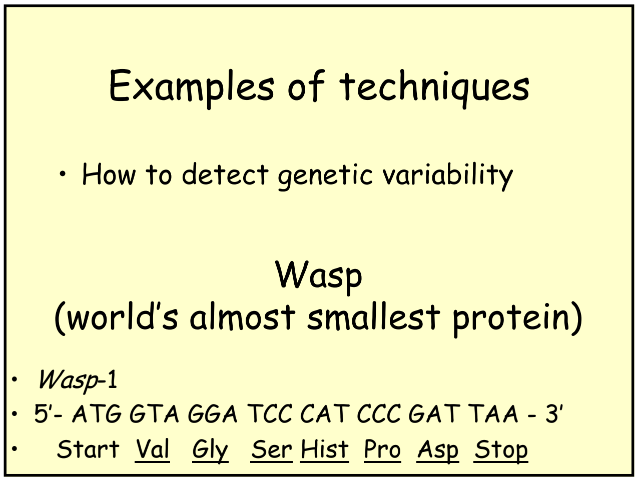
1995-present

- Costs of sequencing declined. *Taq* polymerase plasmids become available….cost of PCR dropped.

2000+

- Whole genome projects on line, pyrosequncing of vast amounts of nucleic acids.

Genome chips, etc?

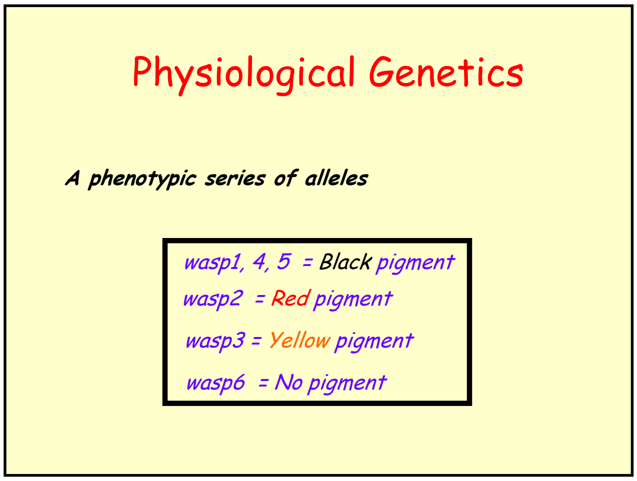


Examples of techniques

- How to detect genetic variability

Wasp
(world's almost smallest protein)

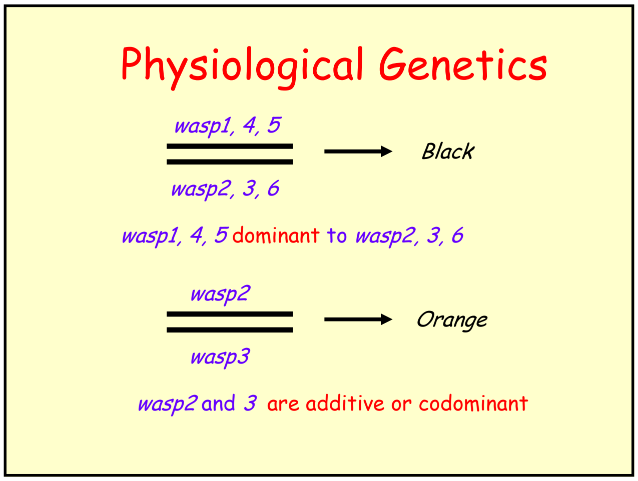
- *Wasp*-1
- 5'- ATG GTA GGA TCC CAT CCC GAT TAA - 3'
- Start <u>Val</u> <u>Gly</u> <u>Ser</u> <u>Hist</u> <u>Pro</u> <u>Asp</u> <u>Stop</u>



Physiological Genetics

A phenotypic series of alleles

wasp1, 4, 5 = Black pigment
wasp2 = Red pigment
wasp3 = Yellow pigment
wasp6 = No pigment



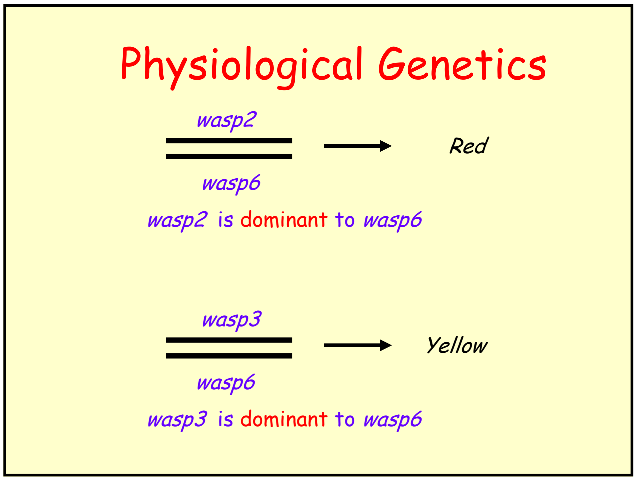
Physiological Genetics

wasp1, 4, 5
wasp2, 3, 6 → *Black*

wasp1, 4, 5 dominant to *wasp2, 3, 6*

wasp2
wasp3 → *Orange*

wasp2 and *3* are additive or codominant



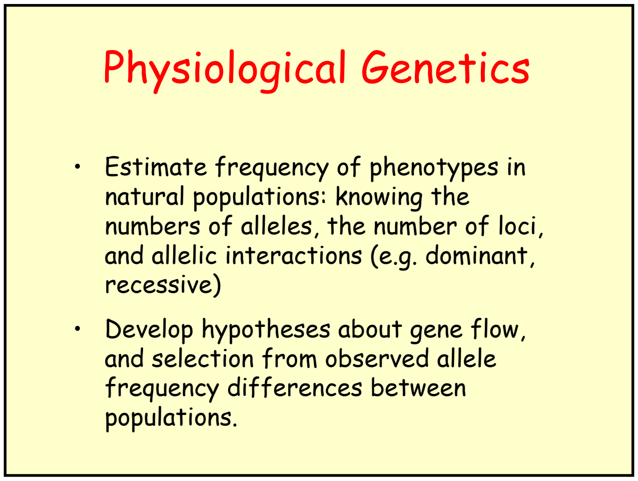
Physiological Genetics

wasp2
wasp6 → *Red*

wasp2 is dominant to *wasp6*

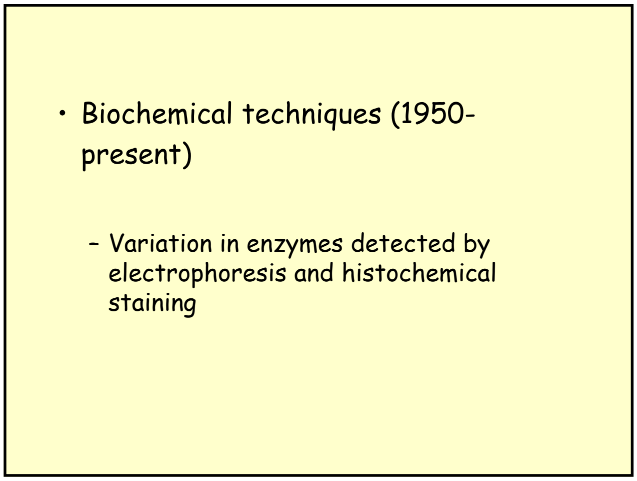
wasp3
wasp6 → *Yellow*

wasp3 is dominant to *wasp6*



Physiological Genetics

- Estimate frequency of phenotypes in natural populations: knowing the numbers of alleles, the number of loci, and allelic interactions (e.g. dominant, recessive)
- Develop hypotheses about gene flow, and selection from observed allele frequency differences between populations.



- Biochemical techniques (1950-present)
  - Variation in enzymes detected by electrophoresis and histochemical staining

Radiation to a germ cell

wasp1

5'- ATG GTA GGA TCC CAT CCC GAT TAA - 3'
Start Val Gly Ser Hist Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 1 (NH_4+) +0 + -1 (COO-) = 0

Adenine misrepaired to a pyrimidine ==> Thymine

wasp2

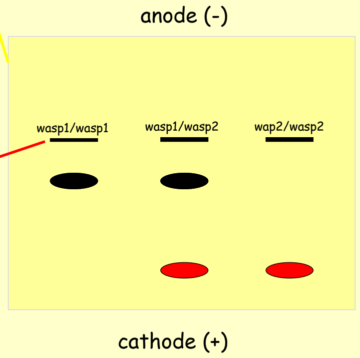
5'- ATG GTA GGA TCC CTT CCC GAT TAA - 3'
Start Val Gly Ser Leu Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 0 + 0 + -1 (COO-) = -1

Allozyme (Isozyme) Electrophoresis

- Supportive media
  - *Starch*
  - *Polyacrylamide*
  - *Agar*
  - *Agarose*
  - *Cellulose acetate*

- Loading slots



Examples from Hedrick (Ch. 1)

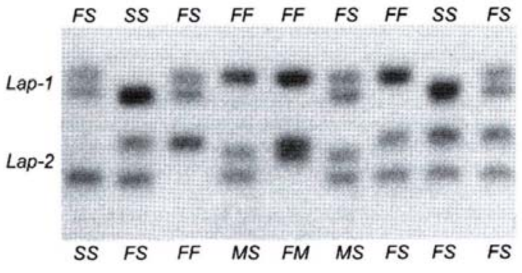


Figure 1.5. Variation in two leucine amino peptidase enzymes in the brown snail, *Helix aspersa* (from Selander, 1976). The upper system (*Lap-1*) is polymorphic for two alleles (*F* and *S*) and the lower system (*Lap-2*) is polymorphic for three alleles (*S*, *M*, and *F*). The genotypes are indicated above and below the gel for the nine individuals pictured.

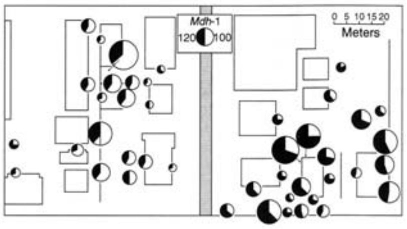


Figure 1.6. The allele frequencies at the *Mdh*-1 locus in brown snail colonies in two city blocks separated by an alley (shaded). Circle size is proportional to colony size, and proportions within the circles indicate allele frequency (from Selander and Kaufman, 1975).

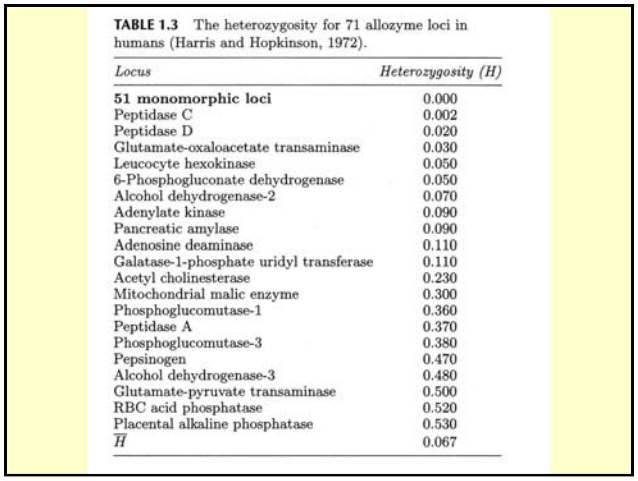
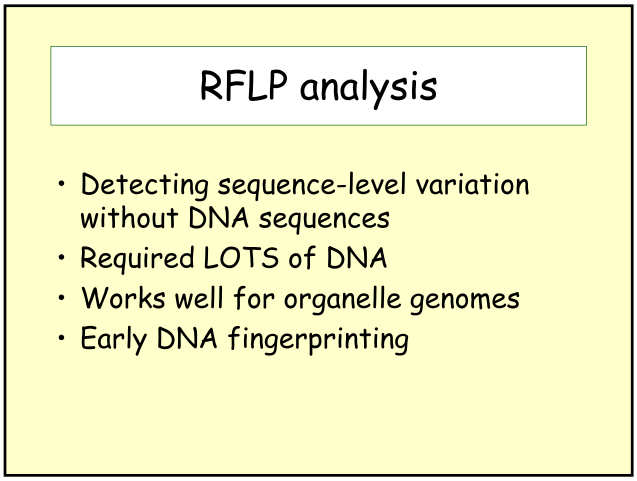


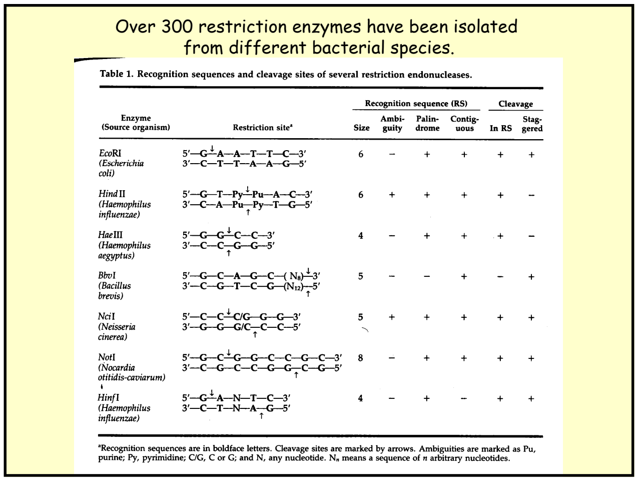
TABLE 1.3 The heterozygosity for 71 allozyme loci in humans (Harris and Hopkinson, 1972).

Locus	*Heterozygosity (H)*
51 monomorphic loci	0.000
Peptidase C	0.002
Peptidase D	0.020
Glutamate-oxaloacetate transaminase	0.030
Leucocyte hexokinase	0.050
6-Phosphogluconate dehydrogenase	0.050
Alcohol dehydrogenase-2	0.070
Adenylate kinase	0.090
Pancreatic amylase	0.090
Adenosine deaminase	0.110
Galatase-1-phosphate uridyl transferase	0.110
Acetyl cholinesterase	0.230
Mitochondrial malic enzyme	0.300
Phosphoglucomutase-1	0.360
Peptidase A	0.370
Phosphoglucomutase-3	0.380
Pepsinogen	0.470
Alcohol dehydrogenase-3	0.480
Glutamate-pyruvate transaminase	0.500
RBC acid phosphatase	0.520
Placental alkaline phosphatase	0.530
$\bar{H}$	0.067



RFLP analysis

- Detecting sequence-level variation without DNA sequences
- Required LOTS of DNA
- Works well for organelle genomes
- Early DNA fingerprinting

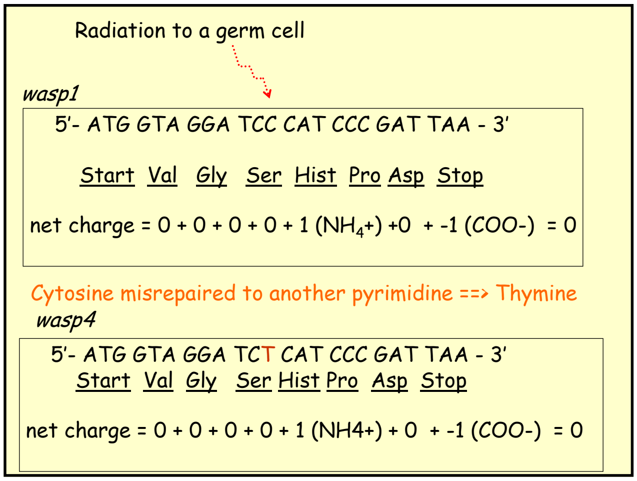


Over 300 restriction enzymes have been isolated from different bacterial species.

Table 1. Recognition sequences and cleavage sites of several restriction endonucleases.

Enzyme (Source organism)	Restriction site[a]	Recognition sequence (RS)				Cleavage	
		Size	Ambiguity	Palindrome	Contiguous	In RS	Staggered
*Eco*RI (*Escherichia coli)*	5′—G↓A—A—T—T—C—3′ 3′—C—T—T—A—A—G—5′	6	−	+	+	+	+
Hind II (*Haemophilus influenzae)*	5′—G—T—Py↓Pu—A—C—3′ 3′—C—A—Pu↑Py—T—G—5′	6	+	+	+	+	−
*Hae*III (*Haemophilus aegyptus)*	5′—G—G↓C—C—3′ 3′—C—C↑G—G—5′	4	−	+	+	+	−
*Bbv*I (*Bacillus brevis)*	5′—G—C—A—G—C—(N_8)↓3′ 3′—C—G—T—C—G—(N_{12})↑5′	5	−	−	+	−	+
*Nci*I (*Neisseria cinerea)*	5′—C—C↓C/G—G—G—3′ 3′—G—G—G/C↑C—C—5′	5	+	+	+	+	+
*Not*I (*Nocardia otitidis-caviarum)*	5′—G—C↓G—G—C—C—G—C—3′ 3′—C—G—C—C—G—G↑C—G—5′	8	−	+	+	+	+
*Hinf*I (*Haemophilus influenzae)*	5′—G↓A—N—T—C—3′ 3′—C—T—N—A↑G—5′	4	−	+	−	+	+

[a]Recognition sequences are in boldface letters. Cleavage sites are marked by arrows. Ambiguities are marked as Pu, purine; Py, pyrimidine; C/G, C or G; and N, any nucleotide. N_n means a sequence of *n* arbitrary nucleotides.



Radiation to a germ cell

wasp1

5′- ATG GTA GGA TCC CAT CCC GAT TAA - 3′

Start Val Gly Ser Hist Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 1 (NH_4+) +0 + -1 (COO-) = 0

Cytosine misrepaired to another pyrimidine ==> Thymine

wasp4

5′- ATG GTA GGA TCT CAT CCC GAT TAA - 3′

Start Val Gly Ser Hist Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 1 (NH4+) + 0 + -1 (COO-) = 0

This 6 bp palindrome is recognized by the restriction enzyme *Bam*H1

wasp1 5'- ATG GTG GGA TCC CAT CCC GAT TAA - 3'
3'- TAC CAC CCT AGG GTA GGG CTA ATT - 5'

wasp1 5'- ATG GTG <u>GGA TCC</u> CAT CCC GAT TAA - 3'
GGA TCC

6 bp palindrome
sequence 5'->3' = sequence 3'->5'

wasp4 5'- ATG GTG <u>GGA TCT</u> CAT CCC GAT TAA - 3'

6 bp palindrome disrupted by substitution

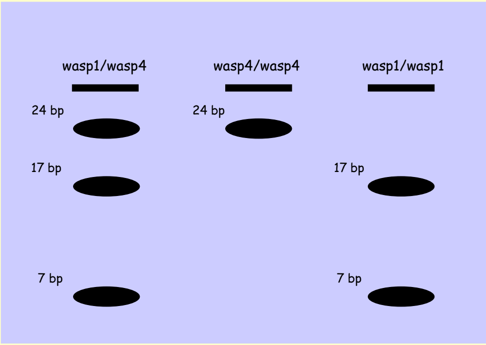
The wasp1 allele is cut by the restriction enzyme *Bam*H1 into a 7 bp fragment with a 5'-GATC-3' and a 17 bp fragment with a 5'-GATC-3' overhang.

7 nucleotides 17 nucleotides

5'- ATG GTG G GA TCC CAT CCC GAT TAA - 3'
3'- TAC CAC CCT AG G GTA GGG CTA ATT - 5'

4 nucleotide overhang

Autoradiograph of a nitrocellulose membrane probed with clones of the wasp gene

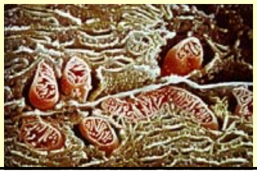


Mitochondrial DNA and RFLP

Analysis of mitochondrial DNA quickly became a powerful tool in the study of animal populations

The mitochondrial genome:

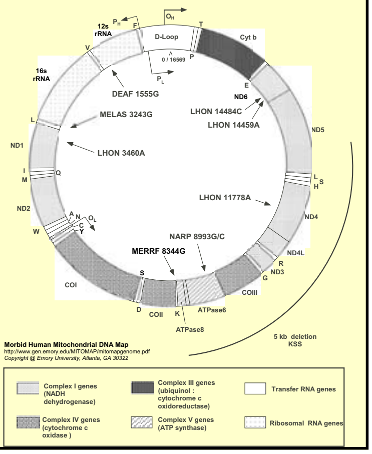
1. Is primarily maternally inherited
2. Does not recombine
3. Evolves at a faster rate than the nuclear genome
4. Intraspecific variation frequently detectable



~ 20 kB in length

Mitochondrial DNA can be used to estimate:

- phylogenetic relationships among maternal lineages
- rates of migration among populations
- genetic diversity among populations



Morbid Human Mitochondrial DNA Map
http://www.gen.emory.edu/MITOMAP/mitomapgenome.pdf
Copyright @ Emory University, Atlanta, GA 30322

RFLP of mitochondrial DNA

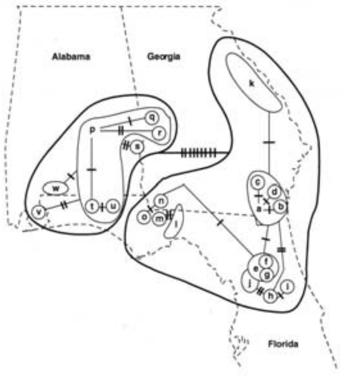


Figure 1.8. The relationship of 23 different mtDNA haplotypes for 87 pocket gophers (from Avise *et al.*, 1979). A network connecting the most related haplotypes is superimposed over the geographic sources of the animals, where the slashes reflect the numbers of inferred differences between haplotypes.

Polymerase Chain Reaction (PCR)

- Mid-1980's
- Amplify a DNA (RNA) fragment starting from only a few originals
  - genetic analysis of any organism, no matter how tiny (e.g. viruses)
  - small amounts of DNA
  - ancient DNA
- Use genomic information from any number of organisms (from GenBank) and apply it to previously unexplored genomes
- Modern genomics meets old school population genetics

Single Strand Conformation Polymorphism (SSCP)

Provides a rapid and inexpensive means to detect 95-99% of all substitutions in an amplified fragment < 500 bp in length

Move toward affordable Single Nucleotide Polymorphisms (SNP's)
Human Hapmap

Radiation to a germ cell

wasp1

5'- ATG GTA GGA TCC CAT CCC GAT TAA - 3'

Start Val Gly Ser Hist Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 1 (NH_4+) +0 + -1 (COO-) = 0

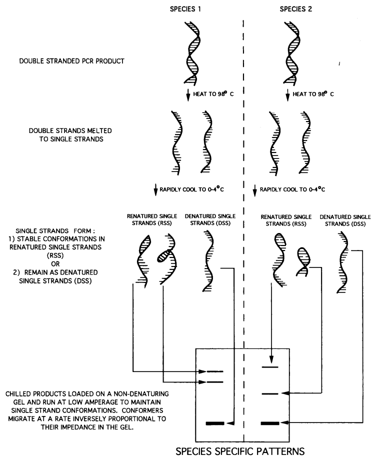
Cytosine misrepaired to another pyrimidine ==> Thymine

wasp4

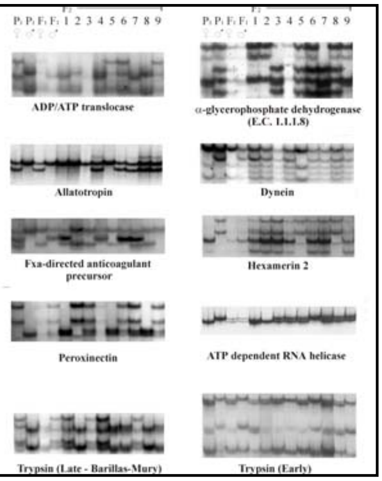
5'- ATG GTA GGA TCT CAT CCC GAT TAA - 3'
Start Val Gly Ser Hist Pro Asp Stop

net charge = 0 + 0 + 0 + 0 + 1 (NH4+) + 0 + -1 (COO-) = 0

Single Strand Conformation Polymorphism (SSCP) analysis


SPECIES 1
SPECIES 2
DOUBLE STRANDED PCR PRODUCT
HEAT TO 98° C
HEAT TO 98° C
DOUBLE STRANDS MELTED TO SINGLE STRANDS
RAPIDLY COOL TO 0-4°C
RAPIDLY COOL TO 0-4°C
RENATURED SINGLE STRANDS (RSS)
DENATURED SINGLE STRANDS (DSS)
RENATURED SINGLE STRANDS (RSS)
DENATURED SINGLE STRANDS (DSS)
SINGLE STRANDS FORM :
1) STABLE CONFORMATIONS IN RENATURED SINGLE STRANDS (RSS)
OR
2) REMAIN AS DENATURED SINGLE STRANDS (DSS)
CHILLED PRODUCTS LOADED ON A NON-DENATURING GEL AND RUN AT LOW AMPERAGE TO MAINTAIN SINGLE STRAND CONFORMATIONS. CONFORMERS MIGRATE AT A RATE INVERSELY PROPORTIONAL TO THEIR IMPEDANCE IN THE GEL.
SPECIES SPECIFIC PATTERNS

SSCP analysis of cDNA markers in *Aedes aegypti*



ADP/ATP translocase

α-glycerophosphate dehydrogenase (E.C. 1.1.1.8)

Allatotropin

Dynein

Fxa-directed anticoagulant precursor

Hexamerin 2

Peroxinectin

ATP dependent RNA helicase

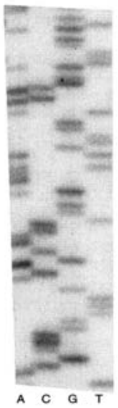
Trypsin (Late - Barillas-Mury)

Trypsin (Early)

Single Nucleotide Polymorphism

- Polymorphism at the nucleotide level
- Within genes:
  - third place synonymous substitution
  - within introns
  - intragenic spacers
- Non-coding DNA
  - both repetitive and non-repetitive DNA
- Detection:
  - RFLP, SSCP
  - DNA sequencing

DNA sequencing: old school versus automated sequencers



Old School

Di-deoxy nucleotides stop PCR reactions and fragment extension

Gel electrophoresis

Figure 1.9. An example of a sequencing gel radiograph, where the different columns indicate the presence of the four nucleotides. The 59-base sequence is from MHC allele *Pooc-6* from the Gila topminnow (Hedrick and Parker, 1998a).

New School: nucleotides labeled with florescent dyes and fragments detected by lasers

Run through capillary tubes

Still depend on PCR and particular primers

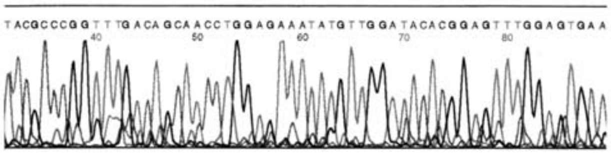


Figure 1.10. An example of the graphical output from an automated sequencer. different positions indicate the presence of different nucleotides (the four different nucleotides from the actual printout are given in different colors). This is the same sequence as given in Figure 1.9 read from bottom to top.

Identification of isozyme (polymorphic enzyme) polymorphism by sequencing of alleles

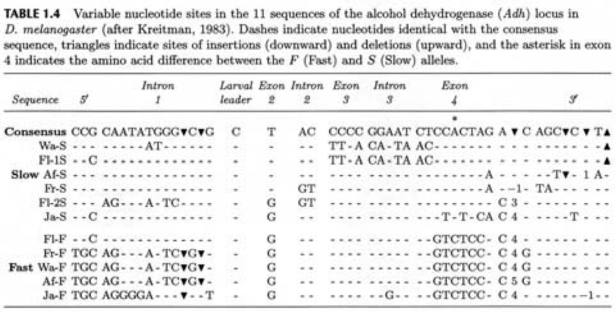


TABLE 1.4 Variable nucleotide sites in the 11 sequences of the alcohol dehydrogenase (*Adh*) locus in *D. melanogaster* (after Kreitman, 1983). Dashes indicate nucleotides identical with the consensus sequence, triangles indicate sites of insertions (downward) and deletions (upward), and the asterisk in exon 4 indicates the amino acid difference between the *F* (Fast) and *S* (Slow) alleles.

Sequence	*5′*	*Intron 1*	*Larval leader*	*Exon 2*	*Intron 2*	*Exon 3*	*Intron 3*	*Exon 4*	*3′*
								*	
Consensus	CCG	CAATATGGG▼C▼G	C	T	AC	CCCC	GGAAT	CTCCACTAG	A ▼ C AGC▼C ▼ T▲
Wa-S	- - -	- - - - - -AT- - - - - - -	-	-	- -	TT-A	CA-TA	AC- - - - - - -	- - - - - - - - - - -▲
Fl-1S	- -C	- - - - - - - - - - - - - -	-	-	- -	TT-A	CA-TA	AC- - - - - - -	- - - - - - - - - - -▲
Slow Af-S	- - -	- - - - - - - - - - - - - -	-	-	- -	- - - -	- - - - -	- - - - - - - -A	- - - - - -T▼- 1 A-
Fr-S	- - -	- - - - - - - - - - - - - -	-	-	GT	- - - -	- - - - -	- - - - - - - -A	- -1- TA- - - - - -
Fl-2S	- - -	AG- - -A-TC- - - -	-	G	GT	- - - -	- - - - -	- - - - - - - - -	C 3 - - - - - - - - - -
Ja-S	- -C	- - - - - - - - - - - - - -	-	G	- -	- - - -	- - - - -	- - -T-T-CA	C 4 - - - - - -T- - -
Fl-F	- -C	- - - - - - - - - - - - - -	-	G	- -	- - - -	- - - - -	- -GTCTCC-	C 4 - - - - - - - - - -
Fr-F	TGC	AG- - -A-TC▼G▼-	-	G	- -	- - - -	- - - - -	- -GTCTCC-	C 4 G - - - - - - - - -
Fast Wa-F	TGC	AG- - -A-TC▼G▼-	-	G	- -	- - - -	- - - - -	- -GTCTCC-	C 4 G - - - - - - - - -
Af-F	TGC	AG- - -A-TC▼G▼-	-	G	- -	- - - -	- - - - -	- -GTCTCC-	C 5 G - - - - - - - - -
Ja-F	TGC	AGGGGA- - -▼- -T	-	G	- -	- - - -	- -G- -	- -GTCTCC-	C 4 - - - - - - - -1- -

Substitutions within exons, insertions and deletions (indels) in introns

SNP’s across a single chromosome

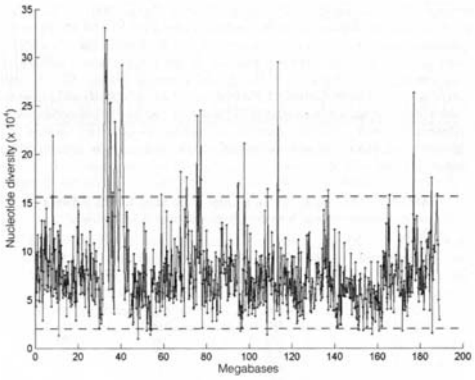


Figure 1.11. The distribution of the nucleotide diversity ($\times 10^4$) across human chromosome 6 as calculated in 200,000-bp bins. The two broken horizontal lines indicate the values within which 95% of the bins fall. The highly variable region around 34 Mb encompasses the 4 Mb of the *HLA* genes.

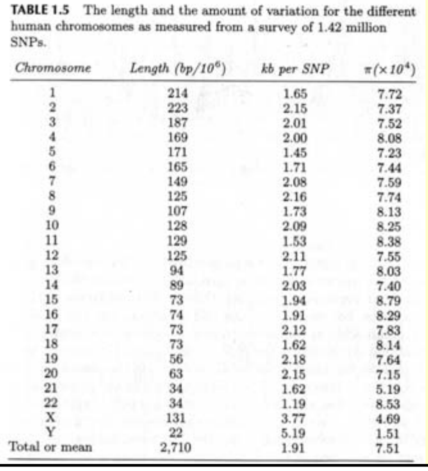


TABLE 1.5 The length and the amount of variation for the different human chromosomes as measured from a survey of 1.42 million SNPs.

Chromosome	*Length ($bp/10^6$)*	*kb per SNP*	*$\pi (\times 10^4)$*
1	214	1.65	7.72
2	223	2.15	7.37
3	187	2.01	7.52
4	169	2.00	8.08
5	171	1.45	7.23
6	165	1.71	7.44
7	149	2.08	7.59
8	125	2.16	7.74
9	107	1.73	8.13
10	128	2.09	8.25
11	129	1.53	8.38
12	125	2.11	7.55
13	94	1.77	8.03
14	89	2.03	7.40
15	73	1.94	8.79
16	74	1.91	8.29
17	73	2.12	7.83
18	73	1.62	8.14
19	56	2.18	7.64
20	63	2.15	7.15
21	34	1.62	5.19
22	34	1.19	8.53
X	131	3.77	4.69
Y	22	5.19	1.51
Total or mean	2,710	1.91	7.51

Microsatellites and VNTR’s

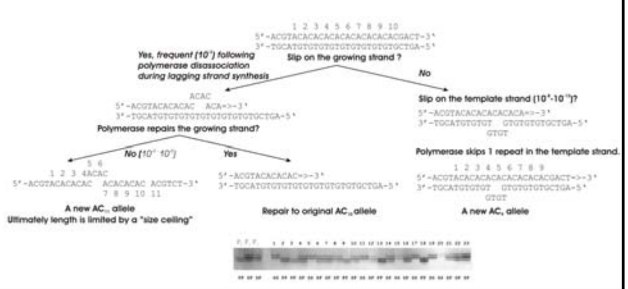
“Microsatellite”: tandem repeat DNA with a unit length of 1-4 bp (Simple Sequence Repeats).

The most common human microsatellites are dinucleotide arrays of $(CA)_n$ which means *n* repeats of CA.

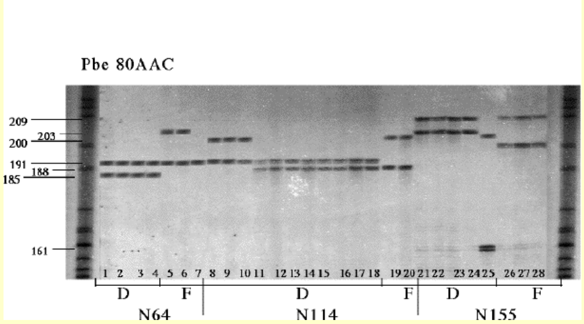
~ 50,000 $(CA)_n$ arrays in the human genome or about one array every 30 kb.

Microsatellites are an abundant component of many (but not ALL genomes).

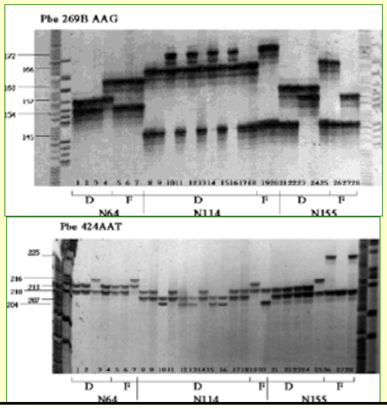
Microsatellites


1 2 3 4 5 6 7 8 9 10
5'-ACGTACACACACACACACACACACGACT-3'
3'-TGCATGTGTGTGTGTGTGTGTGTGTGCTGA-5'
Slip on the growing strand ?
Yes, frequent (10⁻³) following polymerase disassociation during lagging strand synthesis
No
ACAC
5'-ACGTACACACAC ACA=>-3'
3'-TGCATGTGTGTGTGTGTGTGTGTGTGCTGA-5'
Polymerase repairs the growing strand?
No (10⁻¹-10⁻²)
Yes
5 6
1 2 3 4ACAC
5'-ACGTACACACAC ACACACAC ACGTCT-3'
7 8 9 10 11
A new AC₁₁ allele
Ultimately length is limited by a "size ceiling"
5'-ACGTACACACAC=>-3'
3'-TGCATGTGTGTGTGTGTGTGTGTGTGCTGA-5'
Repair to original AC₁₀ allele
Slip on the template strand (10⁻⁸-10⁻¹¹)?
5'-ACGTACACACACACACA=>-3'
3'-TGCATGTGTGT GTGTGTGTGCTGA-5'
GTGT
Polymerase skips 1 repeat in the template strand.
1 2 3 4 5 6 7 8 9
5'-ACGTACACACACACACACACACACGACT=>-3'
3'-TGCATGTGTGT GTGTGTGTGCTGA-5'
GTGT
A new AC₉ allele
P. F. F.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

A denaturing polyacrylamide gel on which microsatellite alleles have been size fractionated and visualized.



Denaturing polyacrylamide gels on which microsatellite alleles have been size fractionated and visualized.

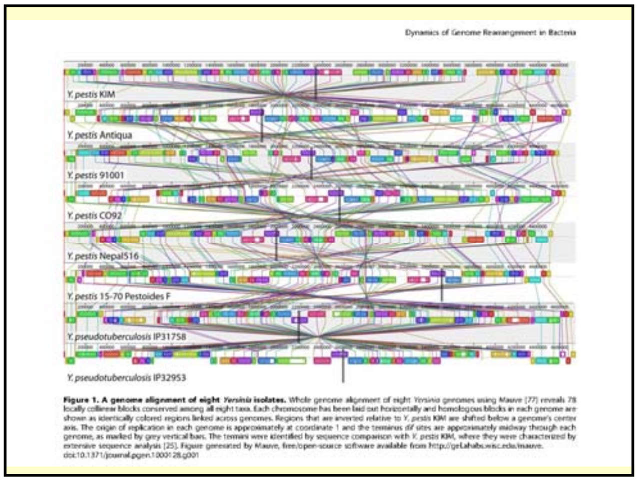


Use of repetitive DNA in population genetics?

- Repetitive elements can diverge in sequence and abundance rapidly
- Potenitally confounds the effects of migration and genetic drift.
- Generally not used in phylogenetics

Aligning Sequences

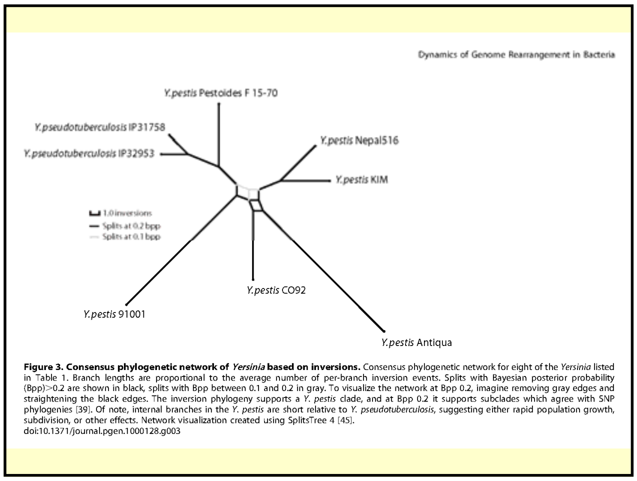
- ClustalW (free on the web, FASTA format)
- BioEdit (free from the web)
- MAUVE (Multiple Genome Alignment: lead time)



Dynamics of Genome Rearrangement in Bacteria

Y. pestis KIM
Y. pestis Antiqua
Y. pestis 91001
Y. pestis CO92
Y. pestis Nepal516
Y. pestis 15-70 Pestoides F
Y. pseudotuberculosis IP31758
Y. pseudotuberculosis IP32953

Figure 1. A genome alignment of eight *Yersinia* isolates. Whole genome alignment of eight *Yersinia* genomes using Mauve [77] reveals 78 locally collinear blocks conserved among all eight taxa. Each chromosome has been laid out horizontally and homologous blocks in each genome are shown as identically colored regions linked across genomes. Regions that are inverted relative to *Y. pestis* KIM are shifted below a genome's center axis. The origin of replication in each genome is approximately at coordinate 1 and the terminus *dif* sites are approximately midway through each genome, as marked by grey vertical bars. The termini were identified by sequence comparison with *Y. pestis* KIM, where they were characterized by extensive sequence analysis [25]. Figure generated by Mauve, free/open-source software available from http://gel.ahabs.wisc.edu/mauve.
doi:10.1371/journal.pgen.1000128.g001



Dynamics of Genome Rearrangement in Bacteria

Figure 3. Consensus phylogenetic network of *Yersinia* based on inversions. Consensus phylogenetic network for eight of the *Yersinia* listed in Table 1. Branch lengths are proportional to the average number of per-branch inversion events. Splits with Bayesian posterior probability (Bpp)>0.2 are shown in black, splits with Bpp between 0.1 and 0.2 in gray. To visualize the network at Bpp 0.2, imagine removing gray edges and straightening the black edges. The inversion phylogeny supports a *Y. pestis* clade, and at Bpp 0.2 it supports subclades which agree with SNP phylogenies [39]. Of note, internal branches in the *Y. pestis* are short relative to *Y. pseudotuberculosis*, suggesting either rapid population growth, subdivision, or other effects. Network visualization created using SplitsTree 4 [45].
doi:10.1371/journal.pgen.1000128.g003