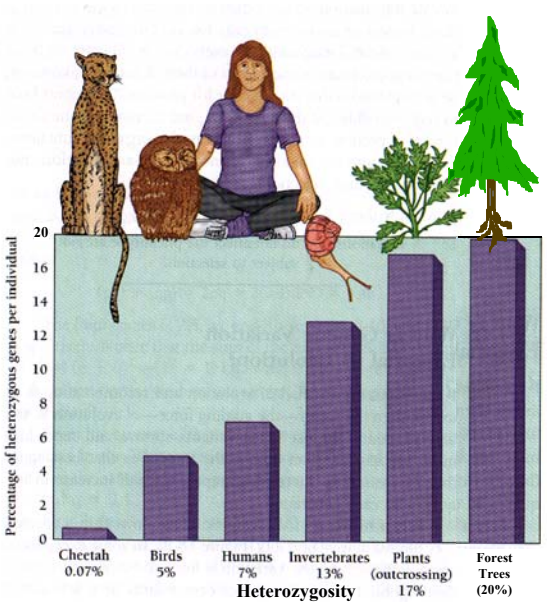
3. Генетический анализ популяций (29 марта, среда):

- measuring of genetic variation at different levels
- genetic variation in individuals, populations and species
- Hardy-Weinberg equilibrium
- genetic analysis of single populations

Levels of genetic variation in Molecular Ecology studies

- **Variation within genes (monomorphic & polymorphic)**
  – alleles & haplotypes
- **Variation within individuals** (no 2 organisms in a sexually reproducing species are the same, except "clones" or monozygotic twins)
  – individual gene or multi-locus haplotype heterozygosity
- **Variation within populations: gene pool**
  – allele frequencies, average heterozygosity, average number of polymorphic alleles and loci and other summary statistics, pairwise individual genetic similarity or distance, sharing indexes
- **Variation between individuals, populations and species**
  – differentiation and genetic distance (pairwise and average)

Levels of genetic variation based on allozyme markers



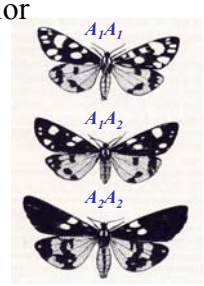
Measuring of genetic variation within and between populations (in space, and how it changes over time)

- allelic and genotypic frequency within populations
- diversity measures (e.g., observed and expected heterozygosities)
- inbreeding index
- migration rate (gene exchange)
- similarity or dissimilarity between populations (allelic and genotypic frequency heterogeneity, distance, differentiation, etc.)

How to estimate allele frequencies

- The scarlet tiger moth (*Panaxia dominula*) has three color forms in England, which differ in their wing coloration.
- This difference is due to two alleles A_1 and A_2 at one diploid locus.
- The following specimen counts were obtained at a particular location in England:

$A_1A_1 = 17062$, $A_1A_2 = 1295$ and $A_2A_2 = 28$.



(Fig. 9.1 in Futuyma, DJ 1998 *Evolutionary Biology*, 3rd ed. Sinauer)

Q1: What are the frequencies of alleles A_1 and A_2?

Total number of moths = **18385**

Total number of alleles = **18385×2 = 36770**

Every A_1A_1 has two A_1 alleles and every A_1A_2 has one A_1 allele

Therefore the number of A_1 alleles is **(2×17062) + (1×1295) = 35419**

The frequency of A_1 (p) is **35419/36770 = 0.963**

Remember $p + q = 1$. Therefore $q = 1 - 0.963 = 0.037$

Q2: Can one predict genotype frequencies, if allele frequencies are known?

Hardy-Weinberg equation


GH Hardy


W Weinberg

- **Godfrey Harold Hardy** (English mathematician) and **Wilhelm Weinberg** (German physician and obstetrician-gynecologist) introduced the principle of population equilibrium in 1908 that is now known as the **Hardy-Weinberg equilibrium**.
  - They suggested equations that describe population equilibrium
  - They described mathematically the relationship between allele frequencies in gene pool and genotype frequencies in population
- After one generation of random mating genotype frequencies will be:

$p^2 + 2pq + q^2 = 1$, which is a binomial expansion of $(p + q)^2 = 1$

A_1A_1 A_1A_2 A_2A_2

where p = frequency of allele A_1, q = frequency of allele A_2 ($p + q = 1$)

- **In an idealized diploid population, the relationship between the allele and the frequencies at a bi-allelic locus are given by the relationship:**

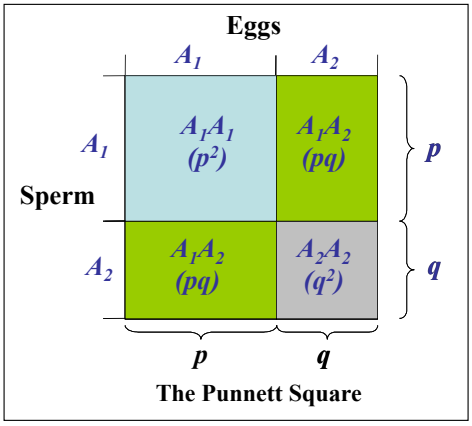
Observed phenotypes	N_P	N_H	N_Q	(N total)
Observed phenotypic frequency	P	H	Q	
Genotypes	A_1A_1	A_1A_2	A_2A_2	
Expected frequency	p^2	$2pq$	q^2	

$$p = \frac{2N_P + N_H}{2N} = \frac{N_P}{N} + \tfrac{1}{2}\frac{N_H}{N} = P + \tfrac{1}{2}H \;,\; q = 1 - p$$

- **There is an equilibrium between observed and expected frequencies**

Hardy-Weinberg Equilibrium

Allele frequencies in one generation can be used to predict genotype frequencies in the next generation



(Reginald C. Punnett 1905)

Frequency of $A_1A_1 = p^2$

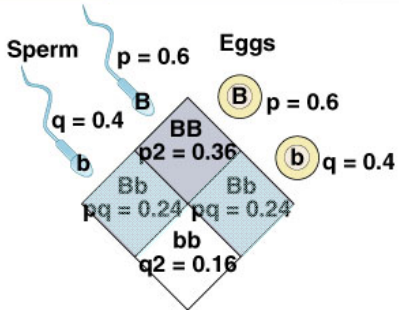
Frequency of $A_1A_2 = 2pq$

Frequency of $A_2A_2 = q^2$

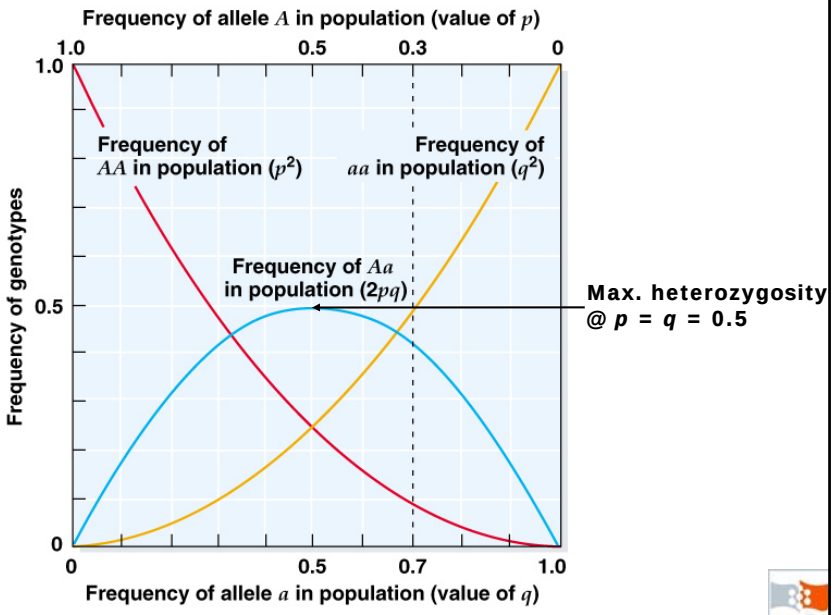
In a random mating population the genotype frequencies are given by: $p^2 + 2pq + q^2 = 1$, which is a binomial expansion $(p + q)^2$

Hardy-Weinberg Equilibrium

Phenotypes			
Genotypes	BB	Bb	bb
Frequency of genotype in population	0.36	0.48	0.16
Frequency of gametes	0.36 + 0.24 = 0.6B		0.24 + 0.16 = 0.4b



Frequencies of genotypes *AA*, *Aa*, and *aa* relative to the frequencies of alleles *A* and *a* in populations at Hardy-Weinberg equilibrium



Hardy-Weinberg Principle

- **Hardy-Weinberg Principle - original proportions of genotypes in a population will remain constant from generation to generation**
  - sexual reproduction (meiosis and fertilization) alone will not change allelic (genotypic) proportions.
  - the Hardy-Weinberg equilibrium allows us to calculate gene frequencies from phenotype frequencies

Hardy-Weinberg Principle: Historical Note

- **Majority of biologists in early 1900's**
  - thought that the dominant allele tends to replace the recessive allele and to become fixed in a population.
  - erroneously believed in blending inheritance
  - were really astonished after G. Hardy & W. Weinberg proved that variation doesn't "disappear" or "blend", and allele frequencies remain constant in large population

Hardy-Weinberg equilibrium

- The Hardy-Weinberg equilibrium allows us to calculate gene frequencies from phenotype frequencies with simplifying assumptions, such as (1) random mating, (2) no selection, (3) large population, (4) no migration, and (5) no mutations.
- It is especially important when one of alleles (*A*) is dominant. Then, frequency (*q*) of recessive allele *a* can be calculated as a square root from the frequency of homozygote *aa* that equals q^2. Then frequency *p* of dominant allele *A* can be calculated as $p = 1 - q$.
- **Example**: Albinism is a rare recessive genetic trait. About one in **20000** individuals is an albino in the US. Therefore, $q^2 = 1/20000$, and thus *q* is approximately **1/141**. According to Hardy-Weinberg equilibrium, one in **141** alleles carries the recessive gene for albinism. How many people are "carriers" or heterozygotes *Aa* for albinism? The probability of being a heterozygote *Aa* is *2pq* which equals **2×(140/141)×(1/141) = 1/70** or about **1.4%**.

Assumptions of Hardy-Weinberg Equilibrium

- **random mating**
- **no selection**
- **a large population size**
- **no migration or gene flow from external sources**
- **no mutation**

(also implicit: diploid organism, sexual reproduction, nonoverlapping generations, equal allele frequencies in the sexes)

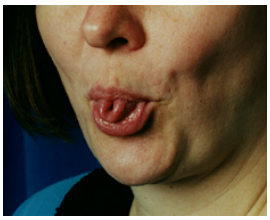
In the real world these assumptions will probably not all be met !

Why then, is HWE useful ?

Hardy-Weinberg Principle: *Why Deal With It?*

- H-W equilibrium *is a starting point (a null hypothesis).* It serves as a model for the genetic structure of a nonevolving population, such as an "ideal" gas in physics.
- It provides a *description of how genetic variation is maintained*.
- *Deviations from H-W may help us learn about evolutionary factors* (*i.e.* selection, migration, mutation, chance, non-random mating).
- We can study what happens when one or more of the assumptions is *violated.* (Although, it is often very difficult to determine causes of deviations from H-W genotypic proportions.)
- *We can use allele rather than genotype frequencies* to study population genetics. There are fewer alleles than genotypes, making calculations easier.

Let's "do" population genetics . . .



Tongue rolling is controlled by 1 locus with 2 alleles, *dominant* (*R*) and *recessive* (*r*)

Phenotype	Genotype
ability	***RR*** or ***Rr***
no ability	***rr***

What is *your* tongue rolling phenotype?

N_{rr} = , $N_{(RR+Rr)}$ = , N =

P_{rr} = , p_r = , q_R = , P_{Rr} =

Let's "do" population genetics . . .

Hitchhiker's thumb is controlled by 1 locus with 2 alleles, *dominant* (*T*) and *recessive* (*t*)

Phenotype	Genotype
straight	***TT*** or ***Tt***
curved	***tt***

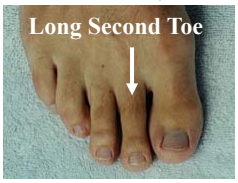
What is *your* hitchhiker's thumb phenotype?

N_{tt} = , $N_{(TT+Tt)}$ = , N =

P_{tt} = , p_t = , q_T = , P_{Tt} =

Let's "do" population genetics . . .

Hallux (big toe on the foot) length is controlled by 1 locus with 2 alleles, *dominant* (*H*) and *recessive* (*h*)



Phenotype (toe length)	Genotype
big toe < 2nd toe	*HH* or *Hh*
big toe ≥ 2nd toe	*hh*

What is *your* hallux phenotype?

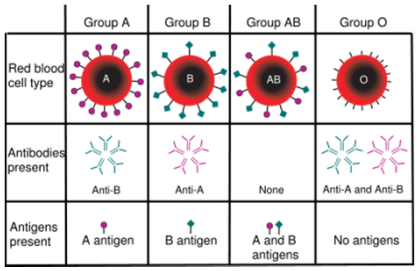
N_{hh} = , $N_{(HH+Hh)}$ = , N =

P_{hh} = , p_h = , q_H = , P_{Hh} =

Let's "do" population genetics . . .

ABO blood type: Controlled by 1 locus with 3 alleles, 2 *dominant* (I^A, I^B) and one *recessive* (*i*)

Genotype	Blood Type Phenotype
ii	O (I)
I^AI^A, I^Ai	A (II)
I^BI^B, I^Bi	B (III)
I^AI^B	AB (IV)



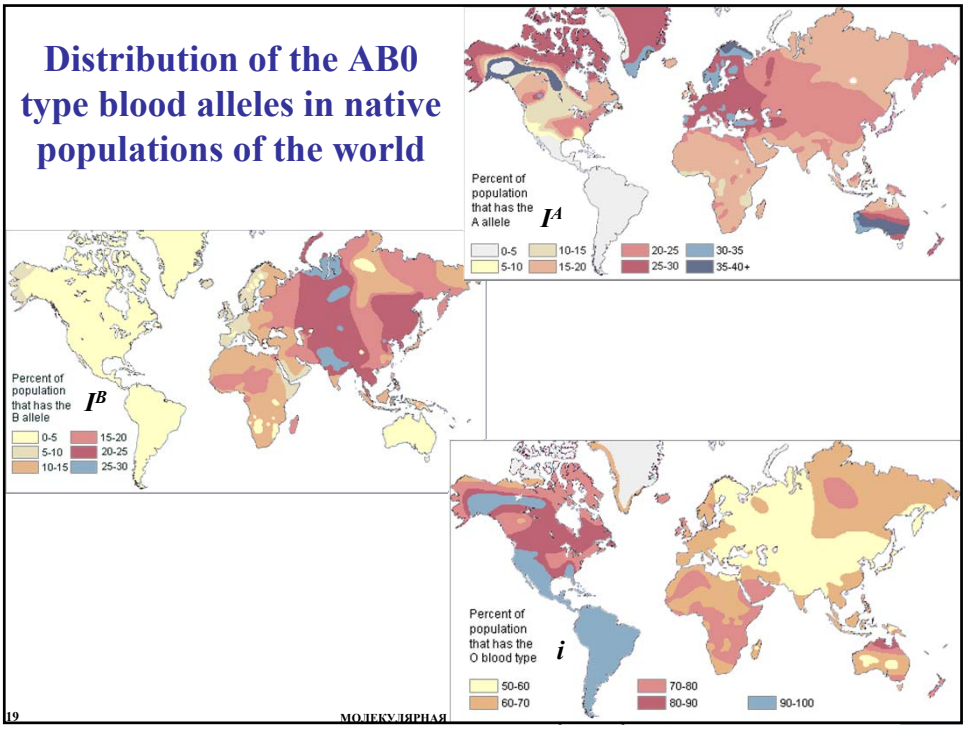
PEOPLE GROUP	O	A	B	AB
Ainu (Japan)	17	32	32	18
Arabs	34	31	29	6
Asian (in USA - General)	40	28	27	5
Blackfoot (N. Am. Indian)	17	82	0	1
Chinese-Canton	46	23	25	6
Chinese-Peking	29	27	32	13
English	47	42	9	3
Hindus (Bombay)	32	29	28	11
Indians (India - General)	37	22	33	7
Indians (USA - General)	79	16	4	1
Japanese	30	38	22	10
Kalmuks	26	23	41	11
Peru (Indians)	100	0	0	0
Russians	33	36	23	8
Spanish	38	47	10	5
Thais	37	22	33	8
USA (blacks)	49	27	20	4
USA (whites)	45	40	11	4

Blood Type Compatibility Chart

Blood Group	Can Recieve From	Can Donate To
A	A·O	A·AB
B	B·O	B·AB
AB	A·B·AB·O	AB
O	O	A·B·AB·O

What is *your* *ABO* blood type?

Distribution of the AB0 type blood alleles in native populations of the world



19 МОЛЕКУЛЯРНАЯ

ABO blood group inheritance

In the United States, in the white population:

45% has blood type ***O:*** **0.45** = genotype ii = expected frequency r^2
41% has blood type ***A:*** **0.41** = genotype I^AI^A and I^Ai = expected frequency $p^2 + 2pr$
10% has blood type ***B:*** **0.10** = genotype I^BI^B and I^Bi = expected frequency $q^2 + 2qr$
4% has blood type ***AB:*** **0.04** = genotype I^AI^B = expected frequency $2pq$

What are frequencies of alleles I^A (p), I^B (q) and i (r)?

$p + q + r = 1$

$p = 1 - q - r = 1-[(q+r)^2]^{½}=1-(q^2+2qr+r^2)^{½} = 1-(B+O)^{½} = 1-(0.10+0.45)^{½} = 0.26$

$q = 1 - p - r = 1-[(p+r)^2]^{½}=1-(p^2+2pr+r^2)^{½} = 1-(A+O)^{½} = 1-(0.41+0.45)^{½} = 0.07$

$r = 1 - p - q = 0.67$

Frequency of I^A allele p = **0.26**
Frequency of I^B allele q = **0.07**
Frequency of i allele r = **0.67**

Mother/Father	*O*	*A*	*B*	*AB*
O	*O*	*O, A*	*O, B*	*A, B*
A	*O, A*	*O, A*	*O, A, B, AB*	*A, B, AB*
B	*O, B*	*O, A, B, AB*	*O, B*	*A, B, AB*
AB	*A, B*	*A, B, AB*	*A, B, AB*	*A, B, AB*

20

ABO blood group inheritance

Blood types observed in a sample of 100 people from a US white population

Blood type	Genotype	Observed	Expected	$\chi^2 = \sum[(\text{Obs-Exp})^2/\text{Exp}]$
O	*ii*	45	$r^2 \times 100 = 44.89$	$(45-44.89)^2/44.89=0.0003$
A	I^AI^A and I^Ai	41	$(p^2 + 2pr) \times 100 = 41.6$	$(41-41.6)^2/41.6=0.0087$
B	I^BI^B and I^Bi	10	$(q^2 + 2qr) \times 100 = 9.87$	$(10-9.87)^2/9.87=0.0017$
AB	I^AI^B	4	$2pq \times 100 = 3.64$	$(4-3.64)^2/3.64=0.0356$
			$\chi^2 = 0.0462$ (< 3.68, *d.f.*=1), $P > 0.05$	

Frequency of *i* allele *r* = **0.67**
Frequency of I^A allele *p* = **0.26**
Frequency of I^B allele *q* = **0.07**

Introduction to some models

- **realistic models are verbal, graphic, or mathematical representations of the real world**
- **deterministic models**

Example 1: B_{t+1} – balance of money next month

$B_{t+1} = B_t + I - E$, where B_t is the balance from this month, and *I* and *E* are income and expenses for the month, respectively (*I* and *E* are called **parameters**)

Example 2: Population growth $R=\frac{N_{t+1}}{N_t}$, where N_t and N_{t+1} are the numbers in generations *t* and *t*+1, respectively. Then, it can be rearranged and $N_{t+1} = R \cdot N_t$. If *R* is the same in other generations then $N_{t+2} = R^2 \cdot N_t$, or in general $N_t = R^t \cdot N_0$

- **stochastic models**

more complicated; often based on the probability distribution; computer simulation based on random numbers (such as a Monte Carlo simulation) can be used to determine the outcome (for example to simulate genetic drift – that is changes in allele frequency due to chance in a finite population

Introduction to probability

- **Правило сложения.** Сумма вероятностей $p_1, p_2, p_3 \ldots p_n$ для всех n возможных взаимоисключающих, несовместимых событий равна их сумме, т.е. единице ($p_1 + p_2 + p_3 \ldots + p_n = 1$).

Example 1: Tossing a coin: p_1 **(head)** $= 0.5$, p_2 **(tail)** $= 1 - p_1 = 0.5$

Example 2: Contracting an infection $p_1 = 0.1$, staying healthy $p_2 = 1 - p_1 = 0.9$

Expected number of outcomes (E) of type 1 for N trials: $E_1 = N \cdot p_1$, the variance is $N \cdot p_i \cdot (1-p_i)$

- **Правило умножения.** Если два события, A и B, независимы (т. е. возникновение одного события не влияет на возможность появления другого), то вероятность того, что оба события произойдут, равна произведению вероятности каждого $P_{(A,B)} = P_A \cdot P_B$. Это правило применимо к любому числу событий.

Example 1: Tossing a coin: The joint probability of obtaining a head from a fair coin four times ($N = 4$) in a row is $P_{(1,1,1,1)} = p_1 \cdot p_1 \cdot p_1 \cdot p_1 = 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.0625$

- In general, if the probability of the first outcome is p and of the second outcome is q, the **binominal probability** that the first type of outcome will occur i times in N independent trials is $P_i = \frac{N!}{i! \cdot j!} p^i \cdot q^j$ (for above example $P_4 = \frac{4!}{4! \cdot 0!} 0.5^4 \cdot 0.5^0 = 0.0625$) **where** ($i + j = N$), or probability of obtaining a head 2 times and a tail 2 times in any order ($N = 4$): $P_4 = \frac{4!}{2! \cdot 2!} 0.5^2 \cdot 0.5^2 = 0.375$)

Introduction to probability

- **If there are more than two possible outcomes, the probability of a combination of outcomes can be calculated from a multinomial probability distribution**

Example 1: Three types of outcomes, such as genotypes A_1A_1, A_1A_2, and A_2A_2. Then, probability P of obtaining i of genotype A_1A_1, j of genotype A_1A_2, and k of genotype A_2A_2 ($i + j + k = N$) is

$$P_{(i,j,k)} = \frac{N!}{i! \cdot j! \cdot k!} \cdot P^i \cdot H^j \cdot Q^k ,$$

where P, H, and Q are the probabilities (frequencies) of the three genotypes in the population from which the sample was taken

($P + H + Q = 1$)

Introduction to probability

Example 1: We know from the previous study that the frequency of the A_1 allele $p = 0.963$, and the A_2 allele $q = 0.037$ in a scarlet tiger moth population

The following three genotypes of scarlet tiger moth individuals were captured in a population: $A_1A_1 = 80$, $A_1A_2 = 15$, and $A_2A_2 = 5$ ($N = 100$)

Then, probability P of obtaining i of genotype A_1A_1, j of genotype A_1A_2 and k of genotype A_2A_2 ($i + j + k = N$) is $P_{(i,j,k)} = \frac{N!}{i!\cdot j!\cdot k!} \cdot P^i \cdot H^j \cdot Q^k$

where P, H, and Q are the probabilities (frequencies) of the three genotypes in the population from which the sample was taken ($P + H + Q = 1$)

Assuming H-W equilibrium $P = p^2$, $H = 2pq$, $Q = q^2$

$$P_{(i,j,k)} = \frac{N!}{i!\cdot j!\cdot k!} \cdot P^i \cdot H^j \cdot Q^k = \frac{N!}{i!\cdot j!\cdot k!} \cdot (p^2)^i \cdot (2pq)^j \cdot (q^2)^k =$$

$$= \frac{100!}{80!\cdot 15!\cdot 5!} \cdot (0.963^2)^{80} \cdot (2\times 0.963\times 0.037)^{15} \cdot (0.037^2)^5 = 5.9e\text{-}10$$

2.3. Exercise & discussion: Answer the questions

DNA, RNA, etc.

1) A nucleic acid was analyzed and found to contain 37% A, 16% G, 22% C, and 25% T.
    1) Does these data present single strand RNA, single strand DNA, double strand RNA, double strand DNA?
    2) What are percentage of purines?
3) Which genomic nucleotide frequency ratios equals to 1 in the following combinations?
    - A/G
    - C/T
    - C/G
    - (A+C)/(G+T)
    - (A+G)/(C+T)
    - (A+T)/(G+C)

Please, fill the blank cells using the genetic code table

	A	C		C							DNA double helix
T			C		T				G	A	
		G						U			mRNA transcribed
											tRNA andticodon
					Trp						Amino acids incorporated into protein

Second Letter

1st letter	U	C	A	G	3rd letter
U	UUU Phe, UUC Phe, UUA Leu, UUG Leu	UCU Ser, UCC Ser, UCA Ser, UCG Ser	UAU Tyr, UAC Tyr, UAA Stop, UAG Stop	UGU Cys, UGC Cys, UGA Stop, UGG Trp	U C A G
C	CUU Leu, CUC Leu, CUA Leu, CUG Leu	CCU Pro, CCC Pro, CCA Pro, CCG Pro	CAU His, CAC His, CAA Gln, CAG Gln	CGU Arg, CGC Arg, CGA Arg, CGG Arg	U C A G
A	AUU Ile, AUC Ile, AUA Ile, AUG Met	ACU Thr, ACC Thr, ACA Thr, ACG Thr	AAU Asn, AAC Asn, AAA Lys, AAG Lys	AGU Ser, AGC Ser, AGA Arg, AGG Arg	U C A G
G	GUU Val, GUC Val, GUA Val, GUG Val	GCU Ala, GCC Ala, GCA Ala, GCG Ala	GAU Asp, GAC Asp, GAA Glu, GAG Glu	GGU Gly, GGC Gly, GGA Gly, GGG Gly	U C A G

Point mutation

- The **5'** **TATTGCACGGTT** **3'** DNA fragment represents an exon
  1) How many amino acids does it encode?
  2) What amino acids does it encode?
- A point mutation substitutes a nucleotide in this sequence - **TATTGCACTGTT**
  3) How many amino acids does it encode?
  4) What amino acids does it encode?

Cell divisions

1) What are the differences between mitosis and meiose?
2) How does meisis explain and confirm the Mendelian law of segregation?
3) Where does independent assortment occur?
4) Why is meiosis needed?
5) Why is it call meiosis (in Greek “reduction”)
6) In a “tetrade”, where are the genes and their alleles?

Recombination

1) What is recombination?
2) Why is it important?
3) Which mendelian principle does it explain and confirm?
4) When are two genes “linked”?
5) Is there a possibility of recombination for linked genes?
6) If yes, how?

Genetic markers

1) What type of genetic markers did Mendel use in his experiments?
2) What were their characteristics?
3) What are allozymes?
4) How can we size DNA fragments of different length?
5) What is a restriction enzyme?
6) Why is PCR useful?

Diversity

1) Why do we care?
2) How can we measure it?
3) How is it created and maintained?

What modes of inheritance do you know?

Breeding dogs

	EB	Eb	eB	eb
EB	EEBB black	EEBb black	EeBB black	EeBb black
Eb	EEBb black	EEbb chocolate	EeBb black	Eebb chocolate
eB	EeBB black	EeBb black	eeBB yellow	eeBb yellow
eb	EeBb black	Eebb chocolate	eeBb yellow	eebb yellow

If in the progeny from crossing two chocolate labradors there were 3 chocolate and 1 gold puppies:

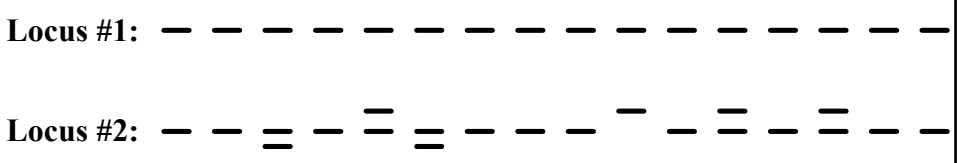
1) What were genotypes of the parents?
2) Is there a chance to have a black puppy out of this cross?

For reference: Estimation of genotype and allele frequencies

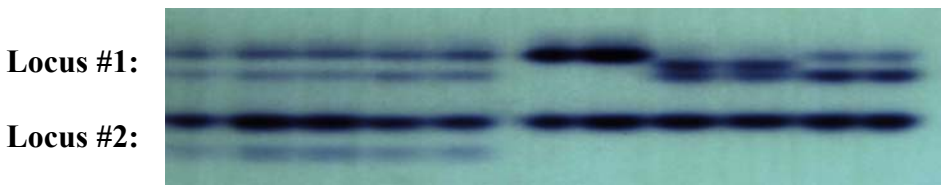
- Frequency of genotype A_xA_y in a population
  - $P_{xy} = N_{xy} / N$
- Frequency of allele A_x
  - $p_x = N_x / 2N$
- Frequency of allele A_x from genotypic frequencies
  - $P_x = P_{xx} + \sum(P_{xy})/2$, for each $x \neq y$

Estimate allele and genotype frequencies for each of 2 loci for each of the following 2 gels

Gel #1:

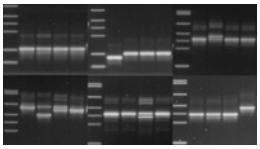


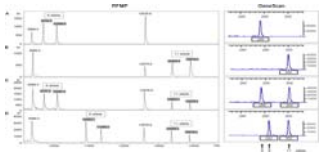
Gel #2:



Enzyme electrophoresis: Codominant marker

Estimate allele and genotype frequencies using the table





Microsatellite SSR:
Codominant marker

1	190	198
2	190	192
3	190	192
4	192	198
5	190	190
6	190	190
7	190	192
8	190	190
9	190	192
10	190	192
11	192	192
12	190	195
13	192	198
14	190	198
15	190	198
16	192	195
17	190	195
18	190	192
19	190	195
20	190	195