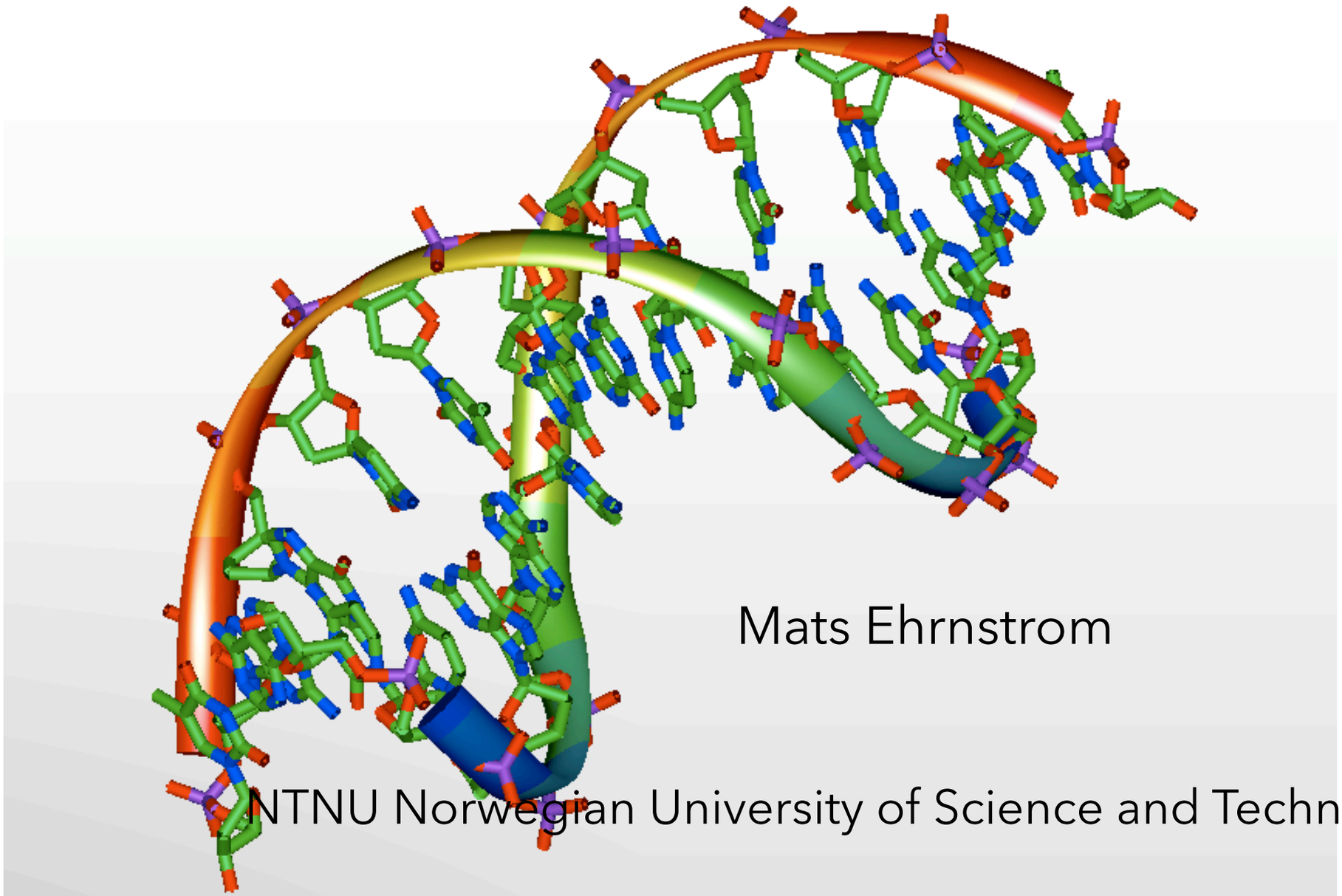
The Hardy-Weinberg Principle and Evolutionary Genetics

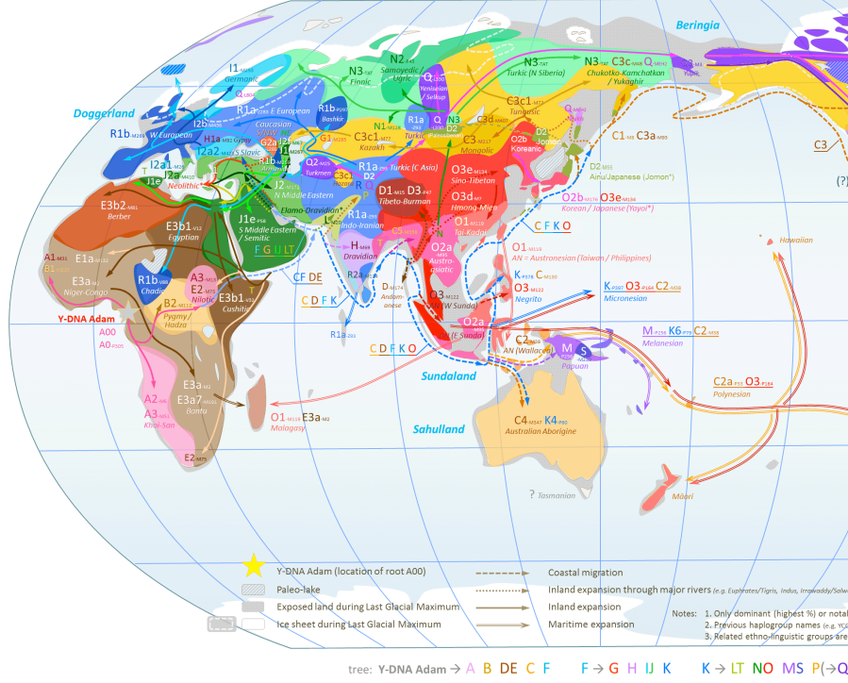


Mats Ehrnstrom

NTNU Norwegian University of Science and Technology

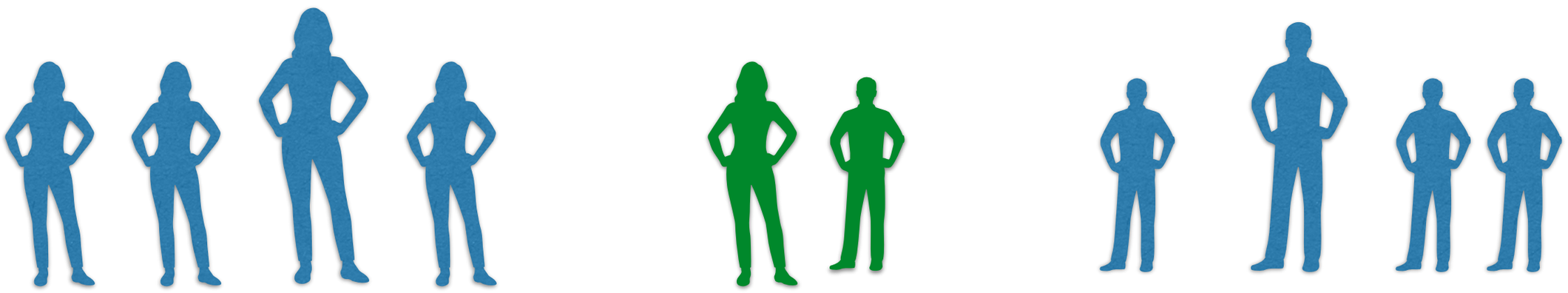
Perleseminaret, December 2021

I come here as an amateur



knowing the exponential function,
multiplication, eigenvalues,
and having the ability to read

– and being interested!



Slide 1: Newton

Newton's first law of motion

s in its state of rest, or in uniform

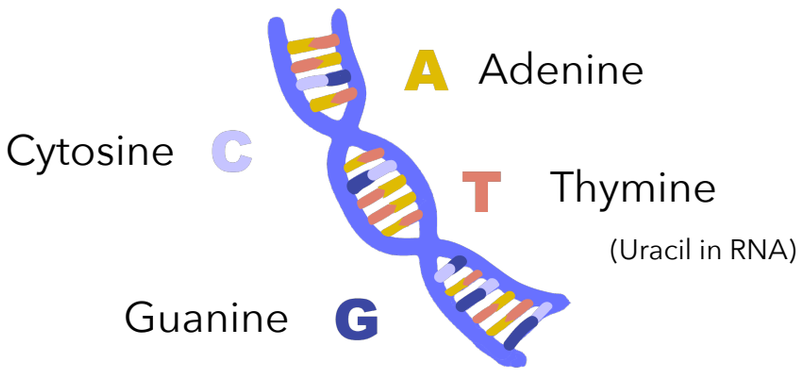
otion in a straight line, unless acted upon by a force.'

Philosophiæ Naturalis Principia Mathematica, 1687

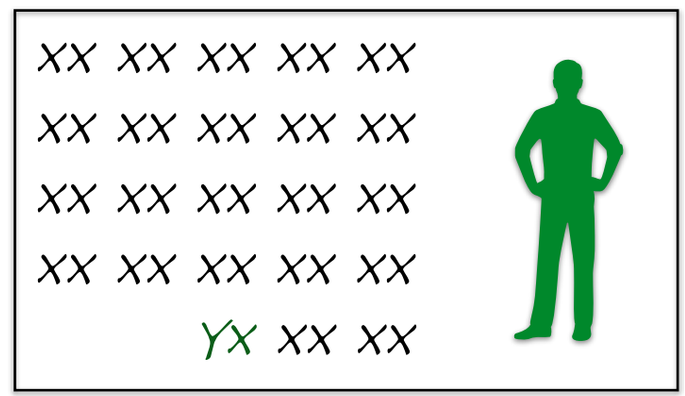
The Hardy–Weinberg principle is the evolutionary genetics equivalent of this law.

The basics and background

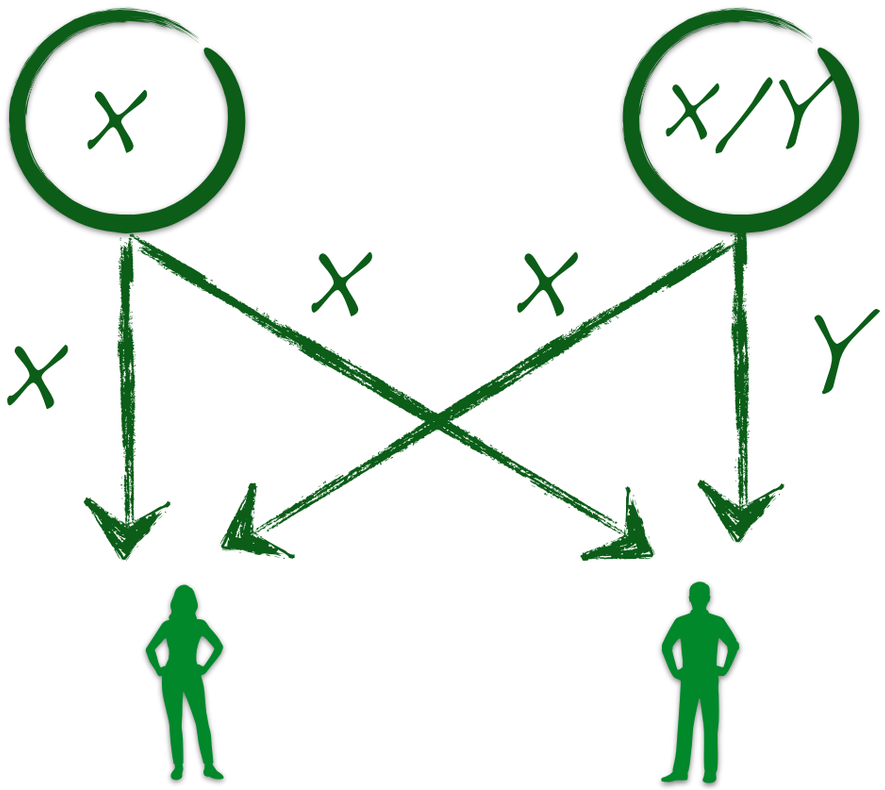
Human cells have 23 chromosome *pairs*

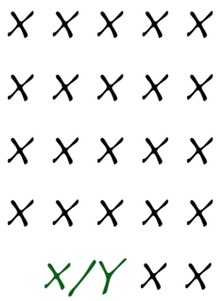


The 23rd pair differs between men and women



Sex cells have 23 *single* chromosomes





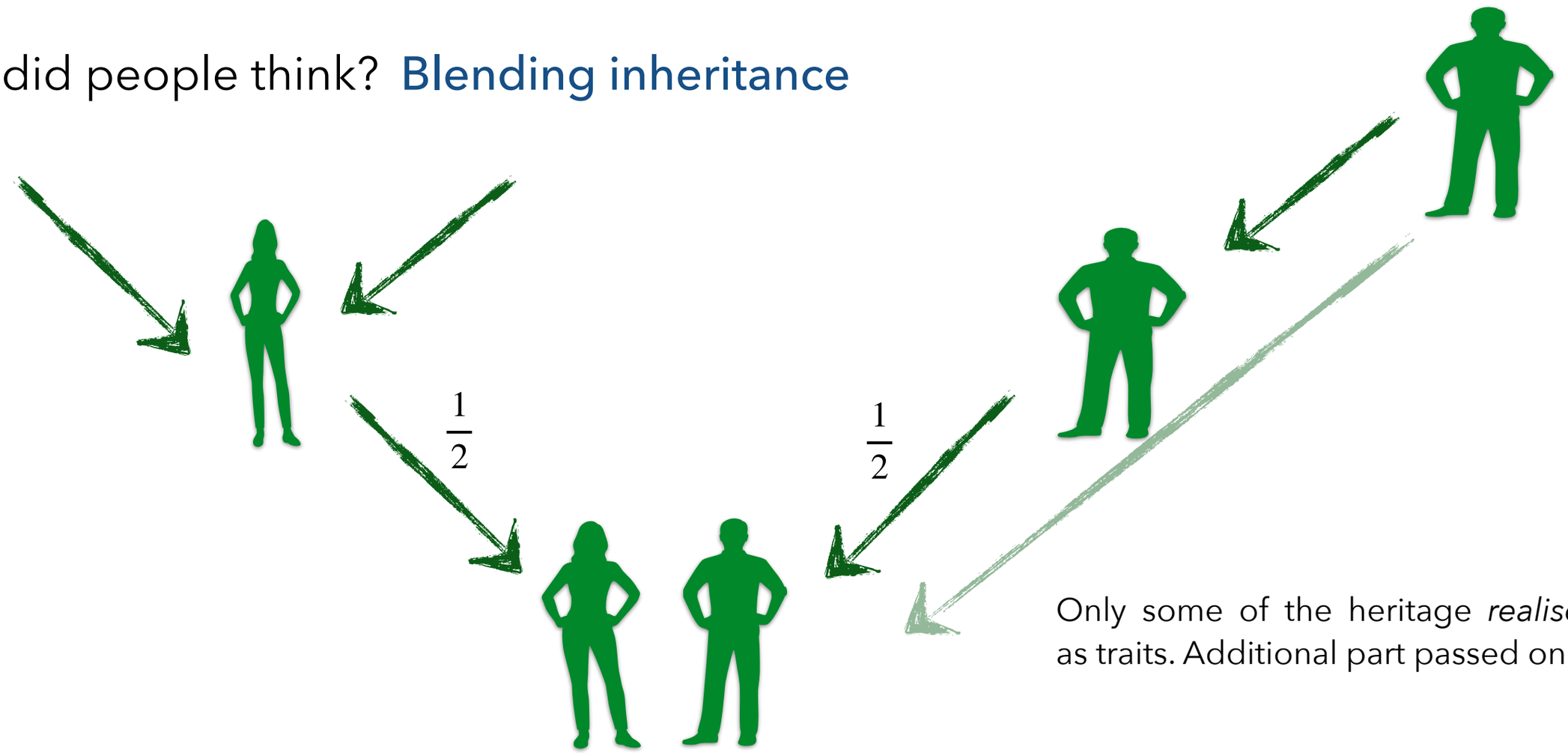
Each male sex cell has *either* an X or Y 23rd chromosome

23 pairs are formed out of these 2 x 23 *single* chromosomes. So the last two chromosomes are inherited with an X from the mother, and either an X or a Y from the father.

But we didn't know this a hundred years ago

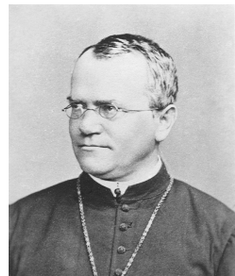
In fact, it was believed we had 48 chromosomes until 1956! (first suggested at all around 1880)

So what did people think? Blending inheritance



$$\frac{1}{2}\text{taller and thinner} + \frac{1}{2}\text{shorter and rounder} = \text{average}$$

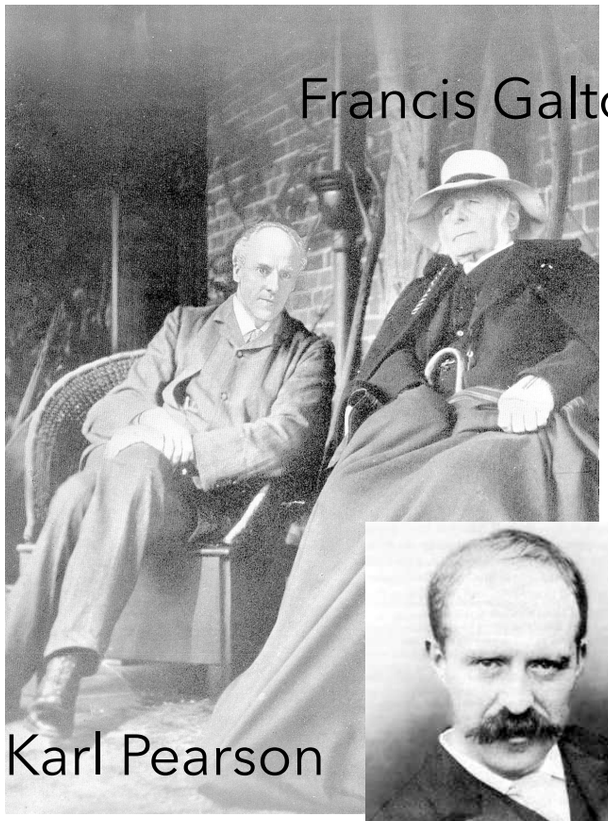
This had been largely recognised for centuries by breeders, and even though Mendel performed his bean experiments in the 1850s and 1860s (*29 000 Pisum Sativum!*) his work was unrecognised until rediscovered around 1900. And then debated for at least 30 years!



The opponents and contributors

1800

1900



Francis Galton

Karl Pearson

Raphael Weldon

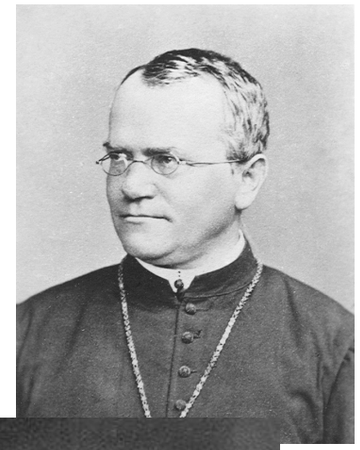
Biometricians

Law of ancestral heredity

Udny Yule

G H Hardy

William Bateson



Gregor Mendel

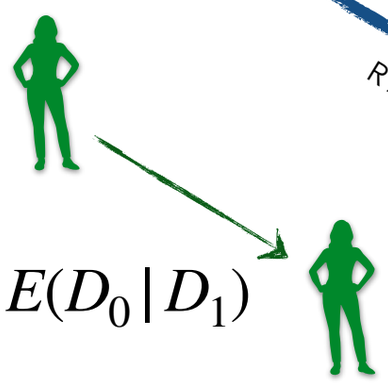
Reginald Punnett

Geneticists

Mendelism

Modern synthesis

R A Fisher, Julian Huxley, …

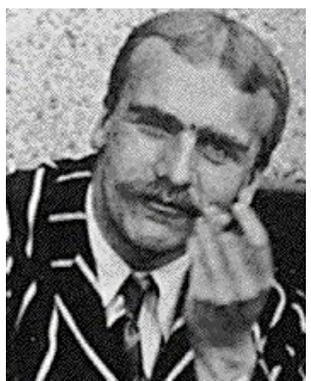


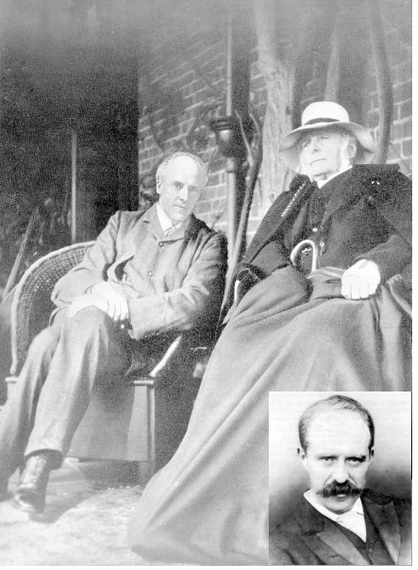
$E(D_0 \mid D_1)$

Modern synthesis

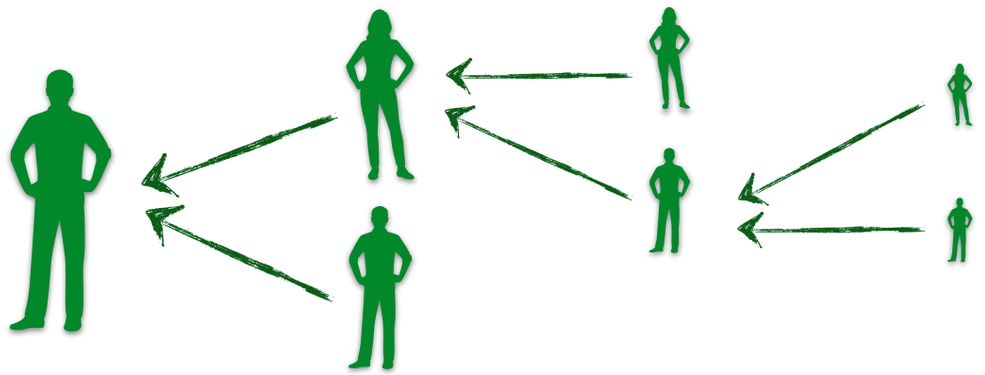
R A Fisher, Julian Huxley, …

J B S Haldane





Galton's law of ancestral heredity

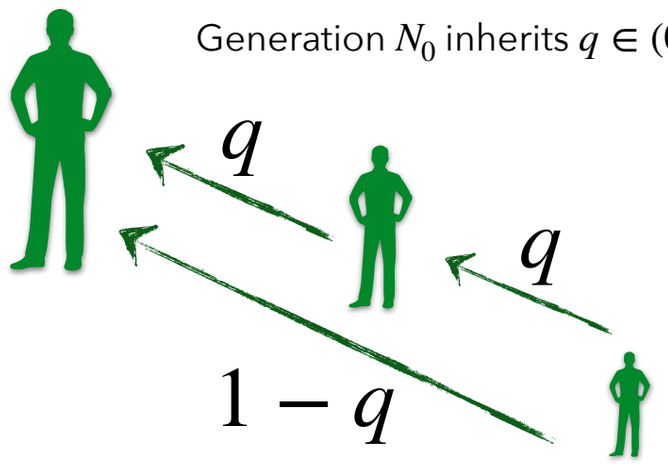


How do the variations in a persons traits depend upon the variations of the same traits in generation n before him?

Galton had many different arguments, but put very roughly they were combinations of (i) empirical data, and (ii) mathematical induction.

From group of persons of the same Stature, to their Kinsmen in various near degrees.	Mean regression=w.	$Q = f = p \times \sqrt{(1 - w^2)}$.
Mid-parents to Sons	2 / 3	1·27
Brothers to Brothers	2 / 3	1·27
Fathers or Sons to Sons or Fathers	1 / 3	1·60
Uncles or Nephews to Nephews or Uncles	2 / 9	1·66
Grandsons to Grandparents...	1 / 9	Practically that of Population, or 1·7 inch.
Cousins to Cousins	2 / 27	

Galton, Natural Inheritance, 1889.



Generation N_0 inherits $q \in (0,1)$ of the traits of Generation N_1.

$$q = \frac{1}{2}$$

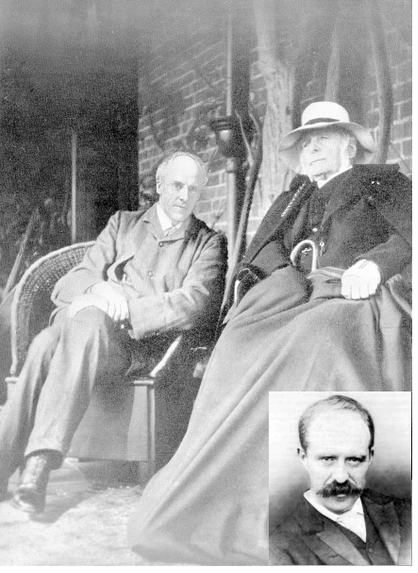
$$1 = \sum_{j=1}^{\infty} q^j = \frac{1}{1-q} - 1$$

all of grandfather's latent and patent traits

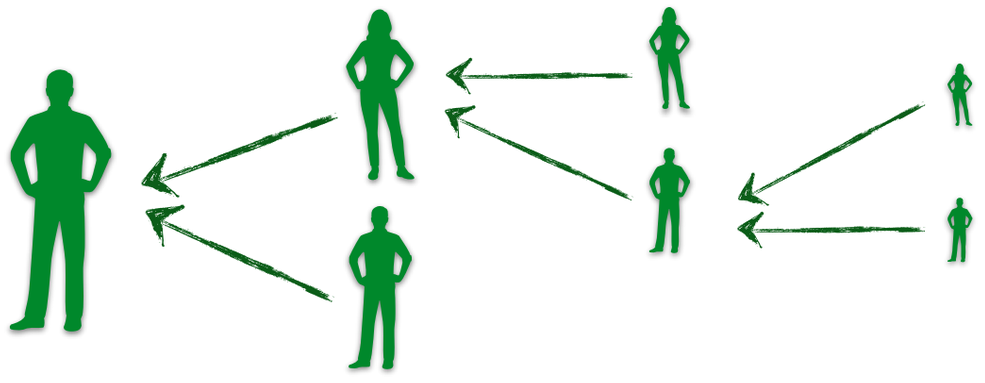
$$1 = q + (1-q) = q + q[q + (1-q)]$$

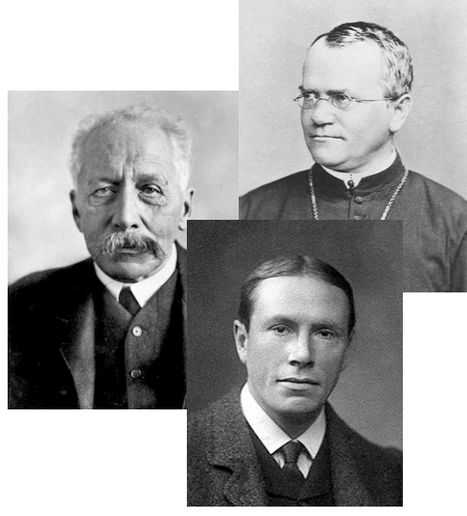
ancestral part — ancestral part

$$= q + q[q + q[q + (1-q)]] =$$



The debate





Simply put, the biometricians tried to solve a multiple regression problem for the different generations by finding a common q, whereas the mendelians choose $q_1 = 1$ and all other $q_j = 0$.

$$x_0 = q_1 x_1 + q_2 x_2 + q_3 x_3 + \ldots$$

> The first question to be asked in such a discussion, is one that does not seem to have occurred to any of Mendel's followers, *viz.:* what, exactly, happens if the two races *A* and *a* are left to themselves to inter-cross freely *as if they were one race?*
>
> Yule, 1902

Yule argued that dominant traits would contribute more than recessive over time, and that Mendel's findings were special cases of the law of ancestral heredity.

Heredity versus equilibrium

Yule

Five hundred dominants whose parents were dominants should therefore produce **425** dominant offspring to **75** recessives, *i.e.*, the chance of their producing a dominant form would be $\frac{425}{500} = \frac{17}{20} = \frac{51}{60}$, while the chance of a dominant of unknown parentage producing a dominant form is only $\frac{50}{60}$. But this is precisely a case of the law of ancestral heredity !

Punnett

Pearson

Pearson had recognised that $\frac{1}{2}, \frac{1}{2}$ for the frequencies of dominant and recessive genes was an equilibrium, so that the frequencies would remain stable in that case.

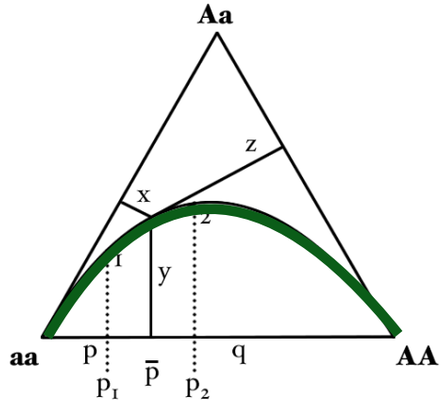
Hardy

Enters Hardy, 1908:

'To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr Udny Yule, to which Mr R C Punnett has called my attention, suggest that it may still be worth making.'

And establishes the Hardy–Weinberg equilibrium

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$$



Hardy's argument

Assume that a gene has two alleles: A (dominant) and a (recessive). Then there are the genotypes AA, Aa and aa.

	A	a
A	AA	Aa
a	aA	aa

Punnett square for two single alleles forming a genotype

Say that in a population the frequencies of genotypes are

AA: p, Aa: $2q$, and aa: r, where $p + 2q + r = 1.$

So when AA mates Aa, there is a 100% chance of picking A from the mother, say, and 50% of picking A (or a) from the father.

In next generation, the frequencies are therefore:

	AA	Aa
AA	p^2	pq
Aa	qp	q^2

Frequencies when AA and Aa mate to AA

AA: $p^2 + 2\frac{1}{2}p(2q) + \left(\frac{1}{2}\right)^2(2q)^2 = p^2 + 2pq + q^2 = (p+q)^2$

AA-AA AA-Aa Aa-Aa

Aa: $2\frac{1}{2}p(2q) + 2\left(\frac{1}{2}\right)^2(2q)^2 + 2pr + 2\frac{1}{2}(2q)r = 2pq + 2q^2 + 2pr + 2qr$

$= 2(q+p)(q+r)$

AA-Aa Aa-Aa AA-aa Aa-aa

aa: $\left(\frac{1}{2}\right)^2(2q)^2 + 2\frac{1}{2}(2q)r + r^2 = q^2 + 2qr + r^2 = (q+r)^2$

Aa-Aa Aa-aa aa-aa

And a check shows that these frequencies sum up to unity!

But that's not the argument

'The interesting question is – in what circumstances will this distribution be the same as that in the generation before?' Hardy 1908

$$p = (p+q)^2 \qquad 2q = 2(q+p)(q+r) \qquad r = (q+r)^2$$

First eq. using $p + 2q + r = 1$:

$$p = p^2 + 2pq + q^2 = p(p+2q) + q^2 = p(1-r) + q^2$$

$$\iff pr = q^2$$

Notice that $pr = q^2$ is fulfilled for example for Pearsons case (all frequencies $\frac{1}{2}$).

First eq. using $p + 2q + r = 1$ and $pr = q^2$:

$$r = q^2 + 2rq + r^2 = q^2 + r(2q+r) = q^2 + r(1-p) = r$$

Third eq. using $p + 2q + r = 1$ and $pr = q^2$:

$$2q = 2q^2 + 2q(r+p) + 2pr$$

$$= 2q^2 + 2q(1-2q) + 2pr = 2q - 2q^2 + 2pr = 2q$$

But that's not the argument

'The interesting question is – in what circumstances will this distribution be the same as that in the generation before?' Hardy 1908

$$p = (p+q)^2 \qquad 2q = 2(q+p)(q+r) \qquad r = (q+r)^2$$

But $\tilde{q}^2 = \tilde{p}\tilde{r}$ will always be the case after one generation, as

$$\tilde{p} = (p+q)^2 \qquad \tilde{q}^2 = (q+p)^2(q+r)^2 \qquad \tilde{r} = (q+r)^2$$

are new the frequencies of the genotypes!

This proves the Hardy–Weinberg equilibrium

$$(p+q)^2 : 2(p+q)(q+r) : (q+r)^2$$

among frequency distributions of the allele pairs AA, Aa and aa,
for *any* frequencies p, q and r.

Assumptions and consequences

Just as in Newton's first law of motion, the assumptions are that there are no exterior forces acting on the population, and that the mating is random process in a population large enough to sustain the probabilities. These are often listed as:

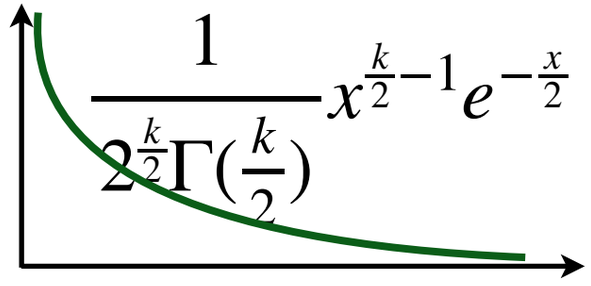
- Infinite population
- random mating
- no natural selection
  (selection would typically favour one genotype over another, and can only decrease genotypes).
- no gene flow or mutations
  (gene flow and mutations would typically increase variance and add genotypes)

The Hardy–Weinberg equilibrium is stable enough to serve as base line for population stratification.

- Given traits in the population one can infer q (but not p directly!) and therefrom calculate p and r using the equilibrium.
- When the mixed genotype Aa is visible as a trait (say grey birds), it is possible to test using Pearson's(!) χ^2-test if the population is within Hardy–Weinberg equilibrium.

$$\chi^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}$$

Goodness of fit for (3 genotypes – 2 alleles) degree of freedom and chosen significance level of the probability distribution given by the density function



So what happens when the gene affects sex cells?

Although not considered by Hardy, consider a gene on the 23rd X chromosome (with no match on the Y chromosome) still with the two alleles A (dominant) and a (recessive).

The female genotypes are as before AA, Aa and aa.

But the male genotypes are only AY and aY.

XX

	AA	**Aa**	**aa**
AY	AA	AA or Aa	Aa
aY	Aa	Aa or aa	aa

Punnett square for female offspring

XY

	AA	**Aa**	**aa**
AY	AY	AY or aY	aY
aY	AY	AY or aY	aY

Punnett square for male offspring

Now we count only **total** frequencies of the alleles A and a in the female and male population, say $p_f + q_f = 1, p_m + q_m = 1.$

$$P_f(AA) = p_f p_m$$
$$P_f(Aa) = p_f q_m + q_f p_m = p_f(1 - p_m) + (1 - p_f)p_m$$
$$P_f(aa) = q_f q_m = (1 - p_f)(1 - p_m)$$

$$P_m(AY) = p_f$$
$$P_m(aY) = 1 - p_f$$

So what happens when the gene affects sex cells?

$$P_f(AA) = p_f p_m$$
$$P_f(Aa) = p_f q_m + q_f p_m = p_f(1-p_m) + (1-p_f)p_m$$
$$P_f(aa) = q_f q_m = (1-p_f)(1-p_m)$$

$$P_m(AY) = p_f$$
$$P_m(aY) = 1 - p_f$$

This gives the following frequencies in the next generation (note that $q = 1 - p$ always here):

$$p_f(t+1) = p_f p_m + \frac{p_f - p_m p_f + p_m - p_m p_f}{2} = \frac{p_f(t) + p_m(t)}{2}$$

$$p_m(t+1) = p_f(t)$$

This is in fact both harder mathematically than the previous case, and more interesting, as we have a second-order recurrence relation for p_f (note that p_m is always chasing p_f):

$$x_{n+2} = \frac{x_{n+1} + x_n}{2}$$

$$x_1 = \tfrac{1}{2}(p_f(0) + p_m(0)), \qquad x(0) = p_f(0)$$

recurrence relations…

First note a few things:

$$x_{n+2} = \frac{x_{n+1} + x_n}{2}$$

Oscillating between sexes

If $p_f(t) > p_m(t)$, then in the next generation,

$$p_f(t+1) = \frac{1}{2}(p_f(t) + p_m(t)) < p_f(t) = p_m(t+1)$$

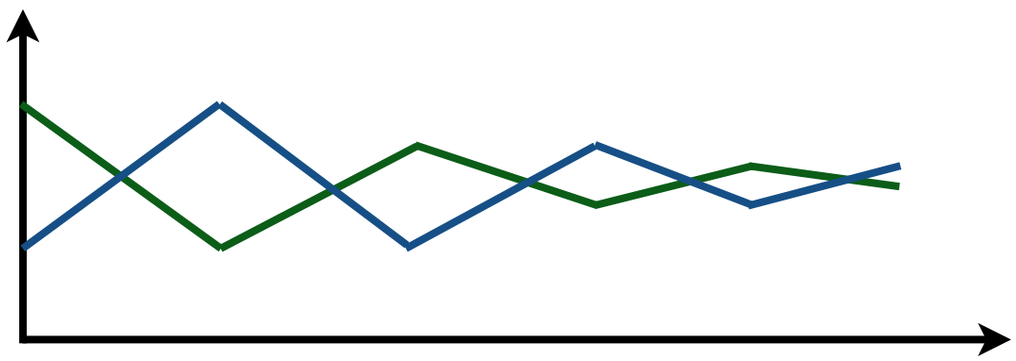
Oscillating in time

If $p_f(t+1) > p_f(t)$, then in the next generation,

$$p_f(t+2) = \frac{1}{2}(p_f(t+1) + p_m(t+1)) = \frac{1}{2}(p_f(t+1) + p_f(t)) < p_f(t+1)$$

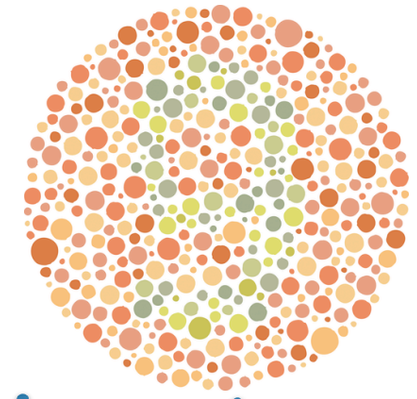
Increasing/decreasing in the subsequences $\{x_n\}_{n=2j}$ and $\{x_n\}_{n=2j+1}$

$$p_f(t+2) = \frac{1}{2}(p_f(t+1) + p_f(t)) \lesseqgtr p_f(t)$$



Always converges – but oscillating around the equilibrium!
(and which equilibrium?)

recurrence relations…



~0.08 ~0.004

Northern Europe

The same allele equilibrium for men and women yields equilibrium phenotypes of frequencies q_{eq} and q_{eq}^2, respectively for men and women, in recessive traits.

Can be cast as a linear algebra problem:

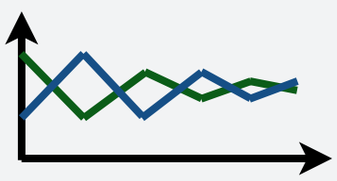
$$\begin{bmatrix} x_{n+2} \\ x_{n+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}$$

Using a change of variables based on eigenvectors/values of the matrix.

$$\begin{bmatrix} y_{n+2} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_{n+1} \\ y_n \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}^n \begin{bmatrix} y_1 \\ y_0 \end{bmatrix}$$

Eigenvalues: $\lambda^2 = \frac{1}{2}(\lambda + 1) \iff (\lambda - \frac{1}{4})^2 = \frac{1+8}{16} \iff \lambda = \frac{1}{4} \pm \frac{3}{4}$

So $x_n = c_1 1^n + c_2 \left(-\frac{1}{2}\right)^n$ with $x_1 = \frac{1}{2}(p_f(0) + p_m(0)), \quad x(0) = p_f(0),$ and we get

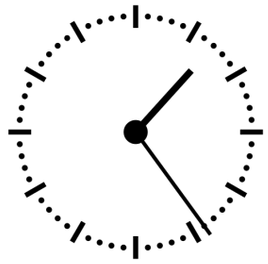


$$x_n = \frac{2}{3}p_f(0) + \frac{1}{3}p_m(0) + \left(p_f(0) - p_m(0)\right)\left(-\frac{1}{2}\right)^n \quad \to \quad \frac{2}{3}p_f(0) + \frac{1}{3}p_m(0)$$

Note that here (most probably) the same $\frac{1}{3}$ fraction as in Galton's tables reappear.

cf. Modern synthesis

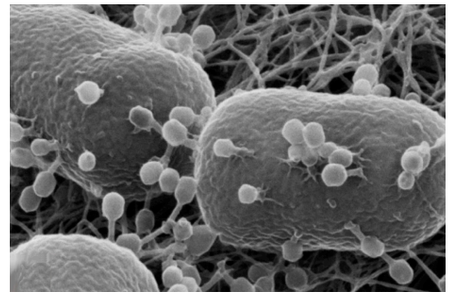
The last slide, and the clock!



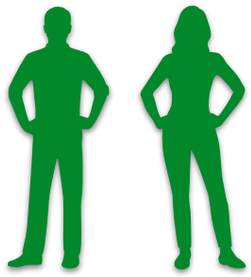
The Hardy–Weinberg equilibrium serves as a 'steady state' to help determine divergence points in evolutionary genetics. The counter-acting mechanism, setting the rate of change in comparison to this steady state, is mutation rate.

1943, the Delbrück–Luria experiment

Nobel laureates Max Delbrück and Salvador Luria used bacteriophages to show that mutations are spontaneous (that is, *prior* to selection) and determined the mutation rate of E coli.



Mutation rates



$\sim 10^{-8}$ per base pair per generation

but higher in sex-specific DNA (Mitochondrial and Y-chromosomal)

$\sim 10^{-11}$

slowest known
Paramecium ciliate

?

$\sim 10^{-3}$

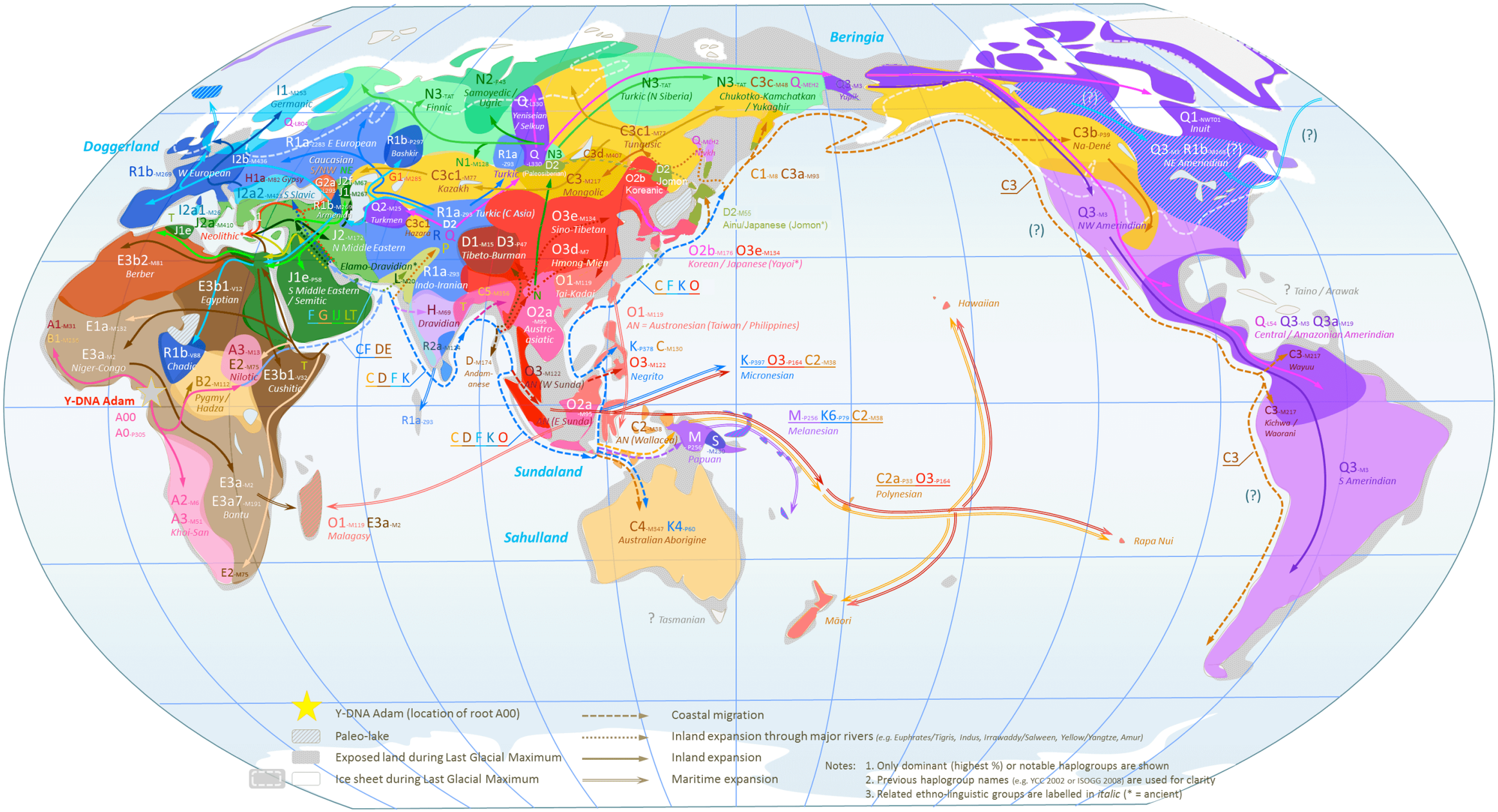
fastest known

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Used differential equations and Poisson distribution test (if mutations would have been a response to the environment, survivors would have distributed according to a Poisson distribution with mean equal variance).

World Map of Y-DNA Haplogroups

Dominant Haplogroups in Native Populations with Possible Migration Routes



tree: **Y-DNA Adam** → A B DE C F F → G H IJ K K → LT NO MS P(→Q R)

Thank you for your attention!