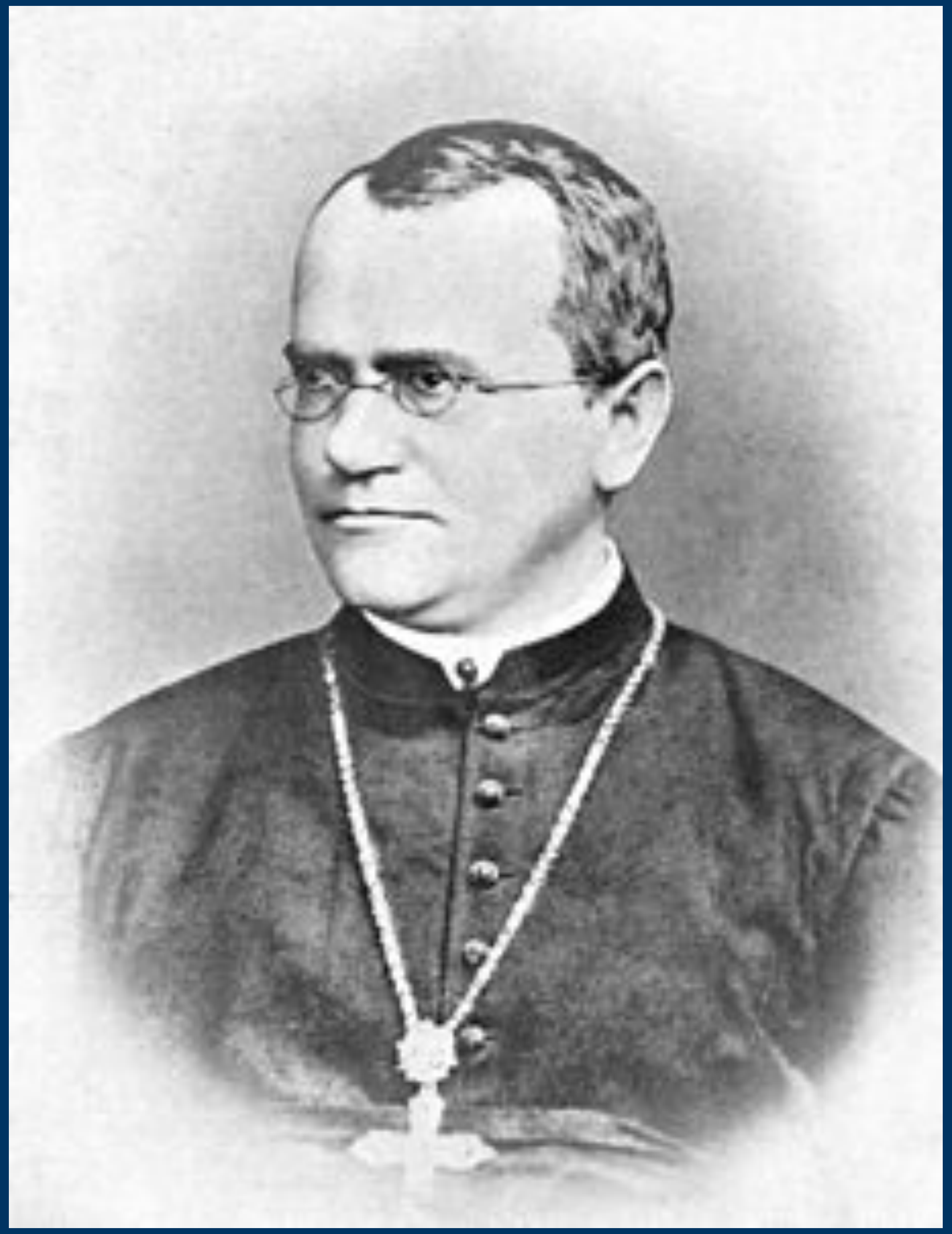
Biostatistics

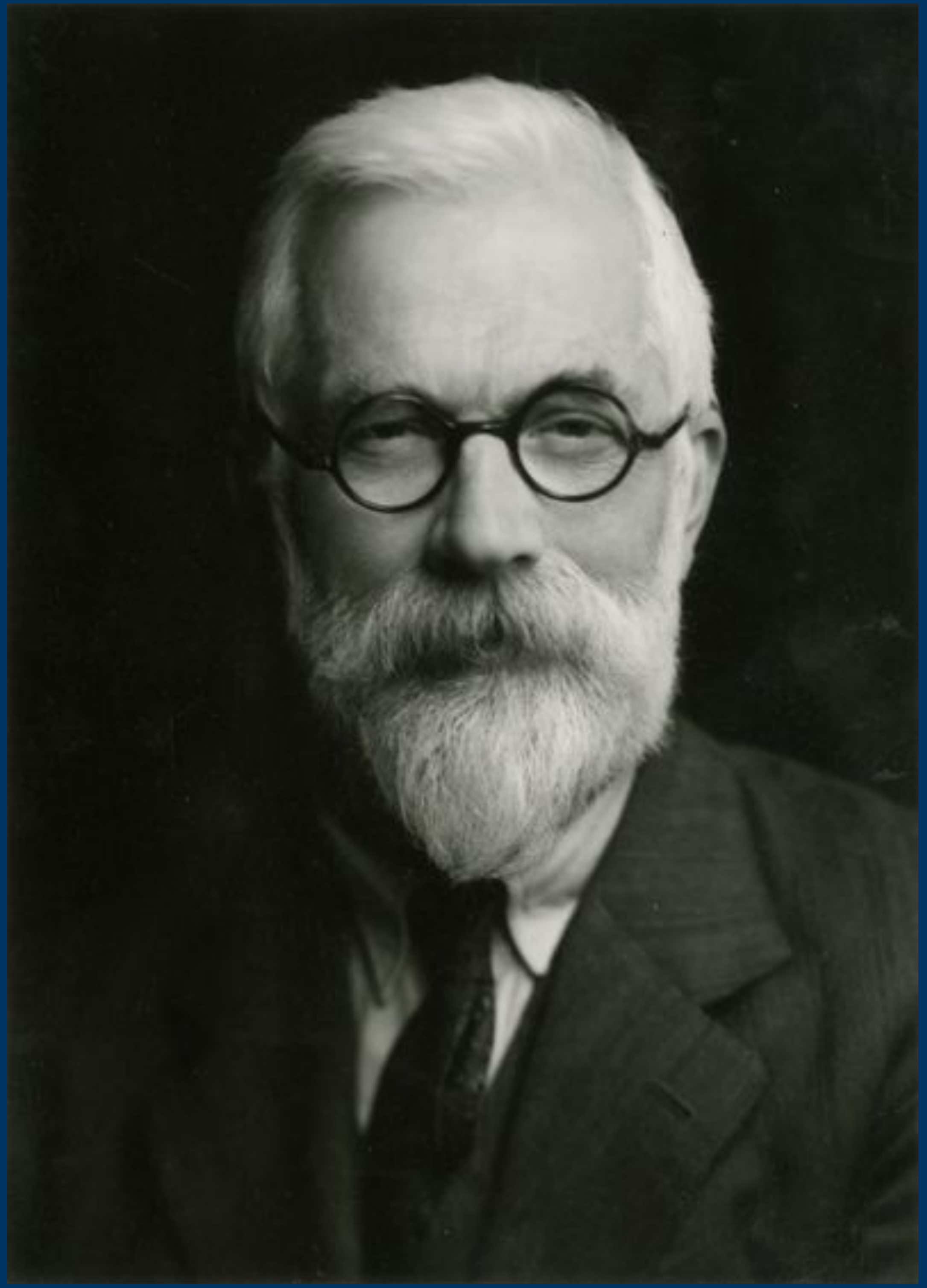
Applications in Genetics and Epigenetics

Nuno Sepúlveda, 01.12.2025

Syllabus

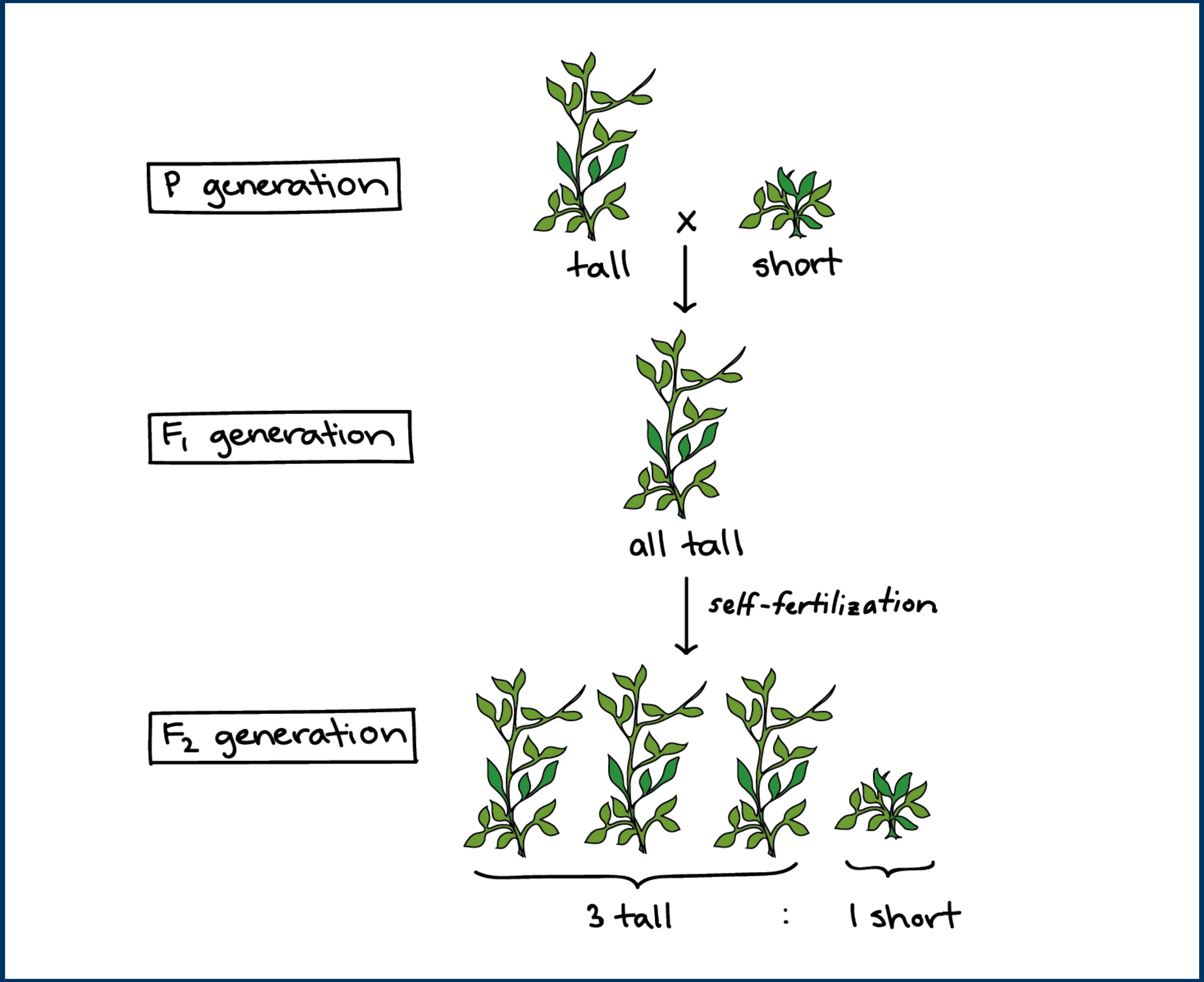
1. **General review**
    a. Population/Sample/Sample size
    b. Type of Data – quantitative and qualitative variables
    c. Common probability distributions/popular tests
2. **Applications in Medicine**
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. **Applications in Genetic and Epigenetic Data**
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values, estimation of biological age
4. **Applications in Serological Data Analysis**
    a. Determination of seropositivity using Gaussian mixture models
    b. Reversible catalytic models for estimating seroconversion rate
    c. Sample size calculation for estimating seroconversion rate





Do you know these people?

Mendelian genetics



Mendelian genetics

Phenotype /Trait = Biological Characteristic Under study (categorical)

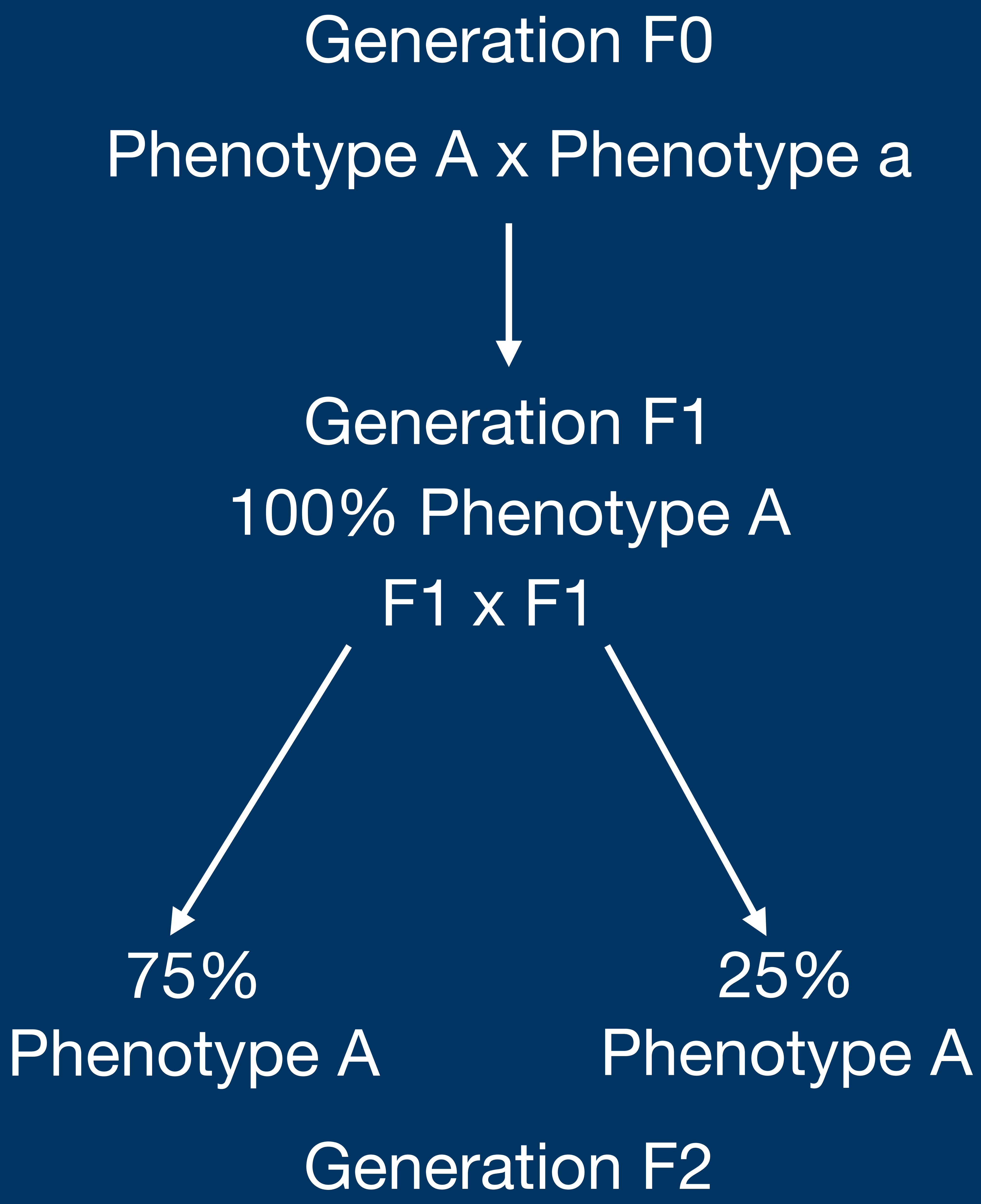
Gene = Unit of Inheritance

Genotype = Composition of gene in terms of alleles

Allele = Variant of a gene

AA

Mendel’s idea/interpretation



Generation F0

Phenotype A x Phenotype a

Generation F1

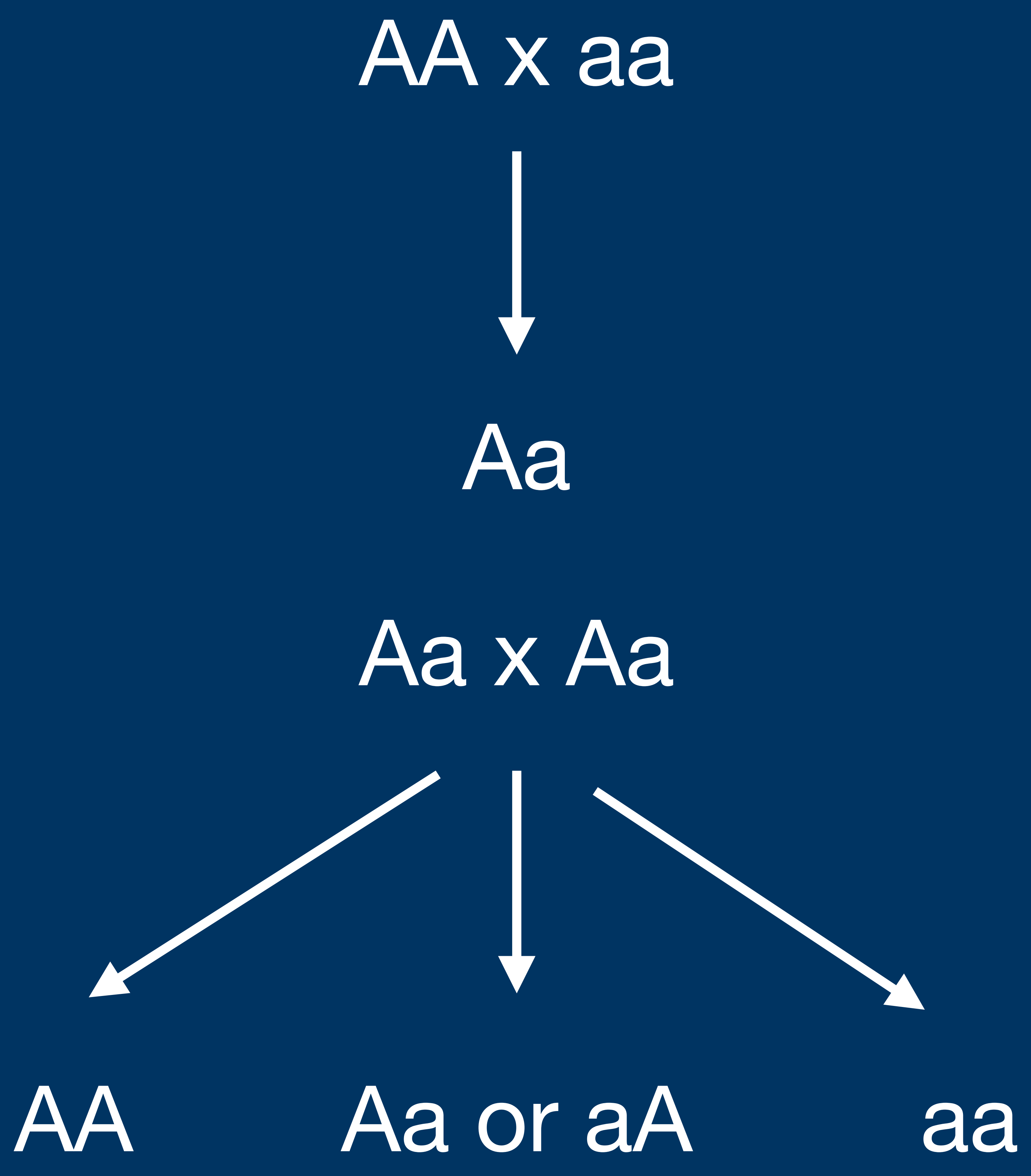
100% Phenotype A

F1 x F1

75%
Phenotype A

25%
Phenotype A

Generation F2



AA x aa

Aa

Aa x Aa

AA

Aa or aA

aa

First two Mendel’s laws

The law of Dominance and Uniformity

Some alleles are dominant over the other alleles for a given gene

The law of Segregation

Two alleles for each gene separate from each other during gametogenesis so that the parent may only pass off one allele; thus, the offspring can only inherit one allele from each parent

Exercise 1: data_mendel_single_trait.csv

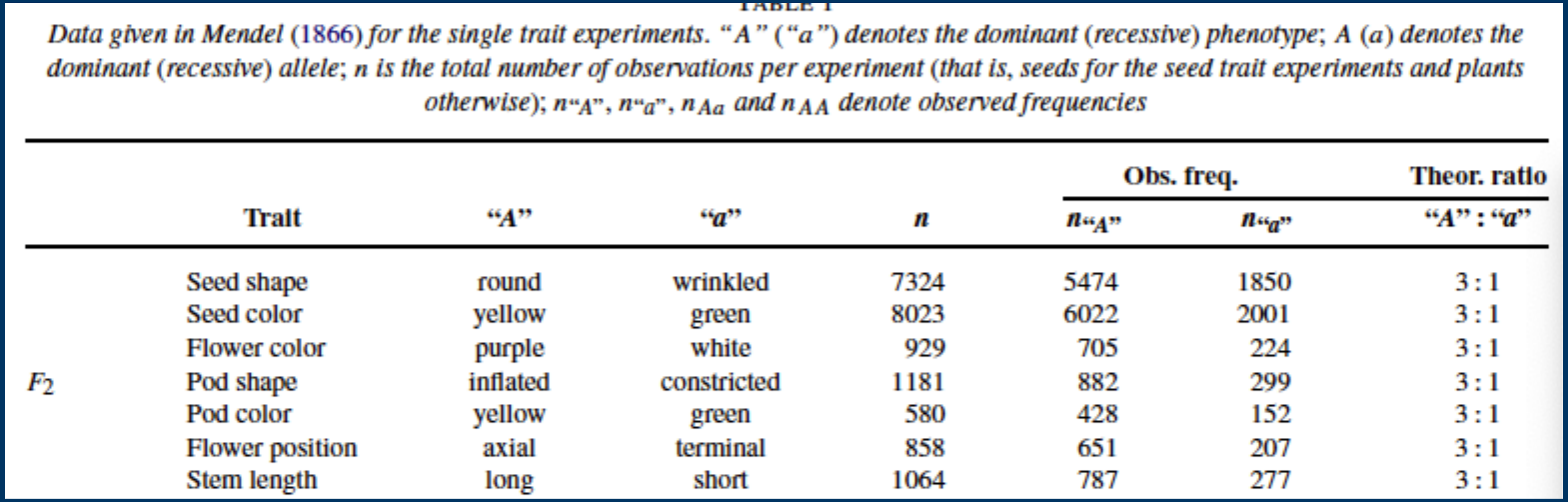


TABLE 1

Data given in Mendel (1866) for the single trait experiments. "A" ("a") denotes the dominant (recessive) phenotype; A (a) denotes the dominant (recessive) allele; n is the total number of observations per experiment (that is, seeds for the seed trait experiments and plants otherwise); $n_{\text{“}A\text{”}}$, $n_{\text{“}a\text{”}}$, n_{Aa} and n_{AA} denote observed frequencies

					Obs. freq.		Theor. ratio
	Trait	"A"	"a"	n	$n_{\text{“}A\text{”}}$	$n_{\text{“}a\text{”}}$	"A" : "a"
	Seed shape	round	wrinkled	7324	5474	1850	3 : 1
	Seed color	yellow	green	8023	6022	2001	3 : 1
	Flower color	purple	white	929	705	224	3 : 1
F_2	Pod shape	inflated	constricted	1181	882	299	3 : 1
	Pod color	yellow	green	580	428	152	3 : 1
	Flower position	axial	terminal	858	651	207	3 : 1
	Stem length	long	short	1064	787	277	3 : 1

Test Mendel's predictions for each trait using an appropriate statistical test.
Draw your conclusions.

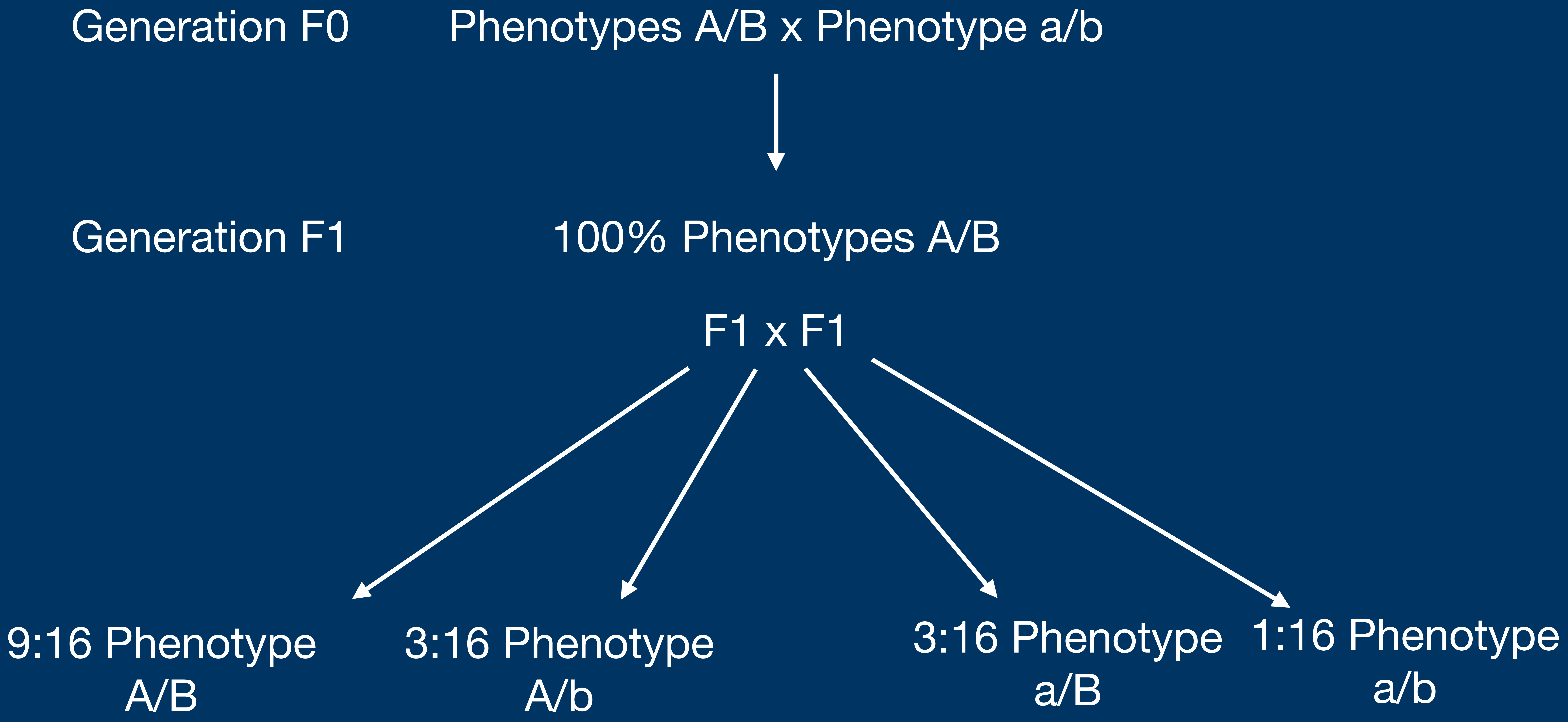
Ana M. Pires. João A. Branco. "A Statistical Model to Explain the Mendel–Fisher Controversy." Statist. Sci. 25 (4) 545 - 565, November 2010.

Third Mendel’s law

The law of Independent Assortment (law of reassortment)

Alleles of different genes segregate independently of one another during gametogenesis

Bifactorial experiments



Generation F0 Phenotypes A/B x Phenotype a/b

Generation F1 100% Phenotypes A/B

F1 x F1

9:16 Phenotype A/B

3:16 Phenotype A/b

3:16 Phenotype a/B

1:16 Phenotype a/b

Bifactorial experiments

Combined genotypes

x	**BB**	**Bb**	**bb**
AA	AA/BB	AA/Bb	AA/bb
Aa	Aa/BB	Aa/Bb	Aa/bb
aa	aa/BB	aa/Bb	aa/bb

Bifactorial experiments

Possibilities (n=16)

Cross	BB	Bb	bb
AA	1	2	1
Aa	2	4	2
aa	1	2	1

Bifactorial experiments

Possibilities (n=16)

Cross	BB	Bb	bb
AA	**Phenotype A/B** 1	2	**Phenotype A/b** 1
Aa	2	4	2
aa	**Phenotype a/B** 1	2	**Phenotype a/b** 1

Bifactorial experiments

Possibilities (n=16)

Cross	**BB**	**Bb**	**bb**
AA	Phenotypes A/B 1	2	Phenotypes A/b 1
Aa	2	4	2
aa	Phenotypes a/B 1	2	Phenotypes a/b 1

Exercise 2:

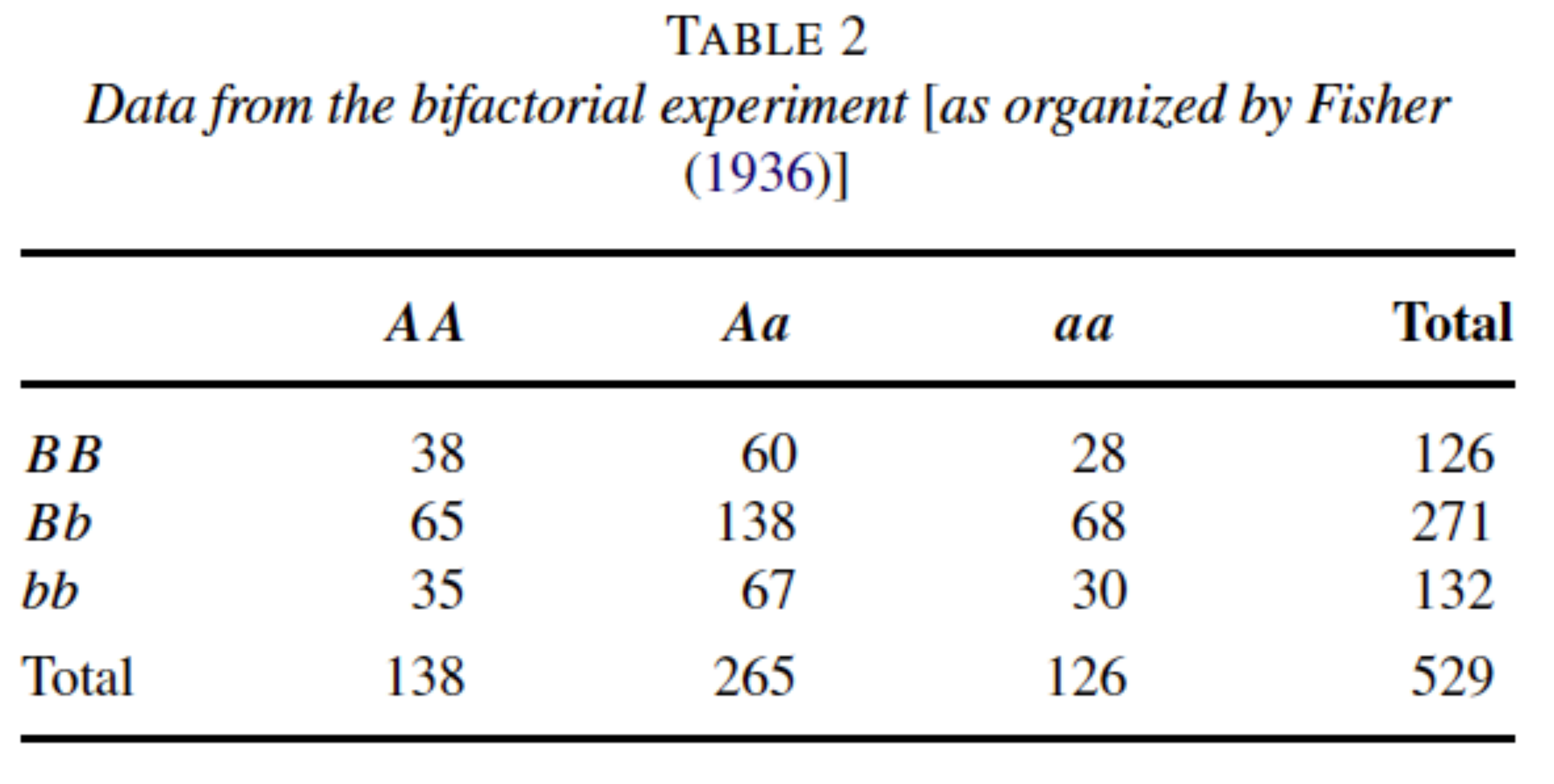


TABLE 2

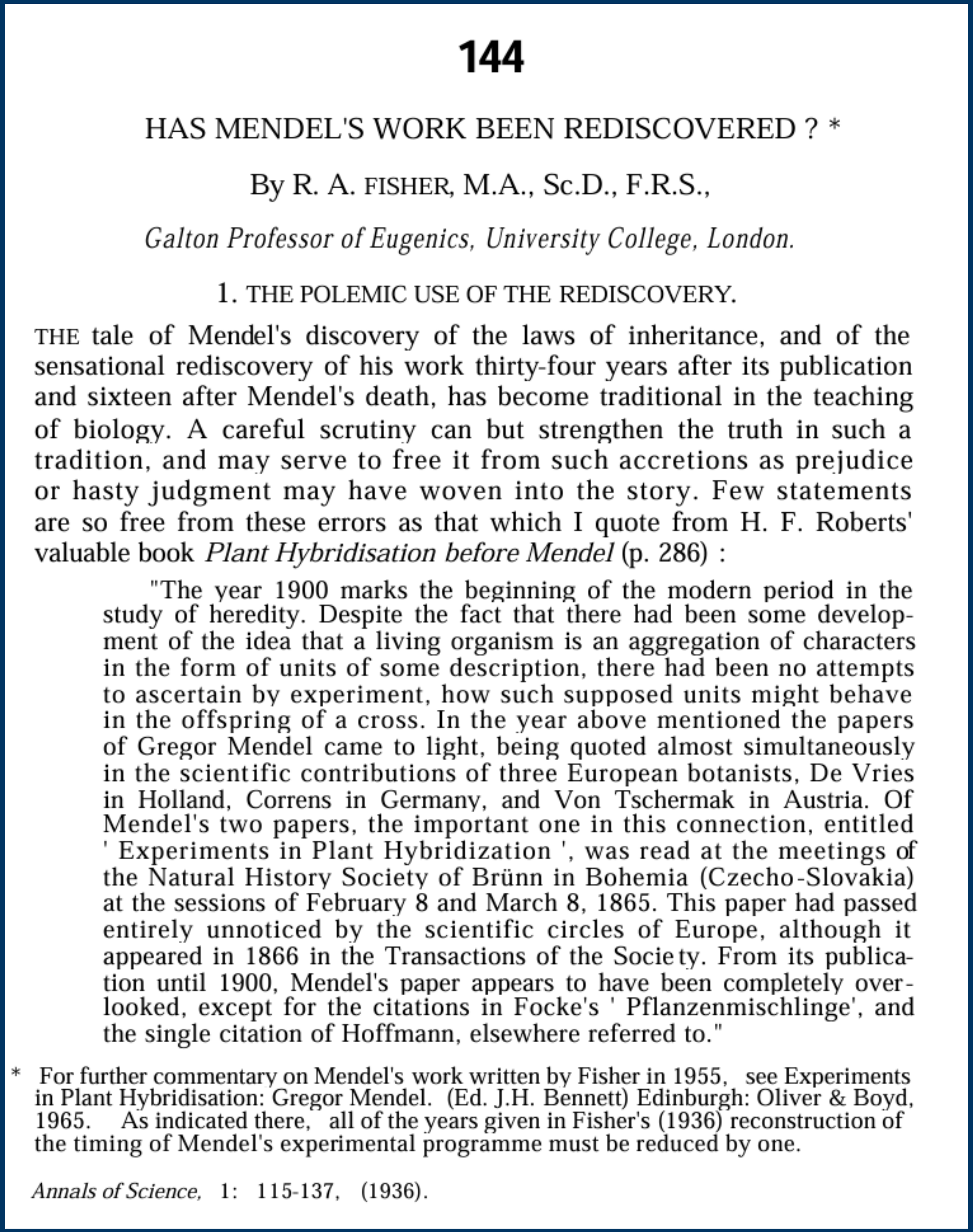
Data from the bifactorial experiment [as organized by Fisher (1936)]

	AA	Aa	aa	**Total**
BB	38	60	28	126
Bb	65	138	68	271
bb	35	67	30	132
Total	138	265	126	529

Test the third Mendel’s law predictions for each trait using an appropriate statistical test.

Draw your conclusions.

Mendel-Fisher Controversy



144

HAS MENDEL'S WORK BEEN REDISCOVERED ? *

By R. A. FISHER, M.A., Sc.D., F.R.S.,

Galton Professor of Eugenics, University College, London.

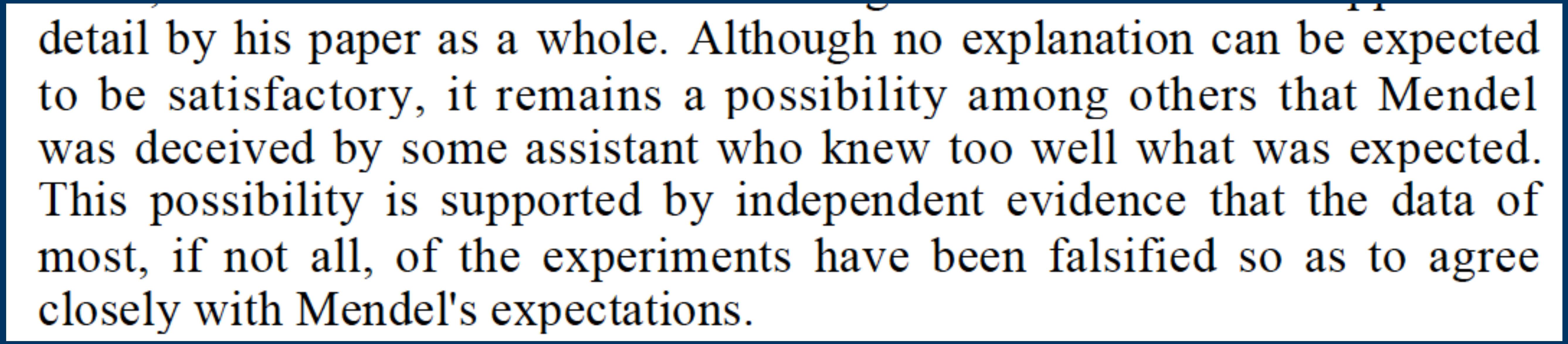
1. THE POLEMIC USE OF THE REDISCOVERY.

THE tale of Mendel's discovery of the laws of inheritance, and of the sensational rediscovery of his work thirty-four years after its publication and sixteen after Mendel's death, has become traditional in the teaching of biology. A careful scrutiny can but strengthen the truth in such a tradition, and may serve to free it from such accretions as prejudice or hasty judgment may have woven into the story. Few statements are so free from these errors as that which I quote from H. F. Roberts' valuable book *Plant Hybridisation before Mendel* (p. 286) :

> "The year 1900 marks the beginning of the modern period in the study of heredity. Despite the fact that there had been some development of the idea that a living organism is an aggregation of characters in the form of units of some description, there had been no attempts to ascertain by experiment, how such supposed units might behave in the offspring of a cross. In the year above mentioned the papers of Gregor Mendel came to light, being quoted almost simultaneously in the scientific contributions of three European botanists, De Vries in Holland, Correns in Germany, and Von Tschermak in Austria. Of Mendel's two papers, the important one in this connection, entitled ' Experiments in Plant Hybridization ', was read at the meetings of the Natural History Society of Brünn in Bohemia (Czecho-Slovakia) at the sessions of February 8 and March 8, 1865. This paper had passed entirely unnoticed by the scientific circles of Europe, although it appeared in 1866 in the Transactions of the Society. From its publication until 1900, Mendel's paper appears to have been completely overlooked, except for the citations in Focke's ' Pflanzenmischlinge', and the single citation of Hoffmann, elsewhere referred to."

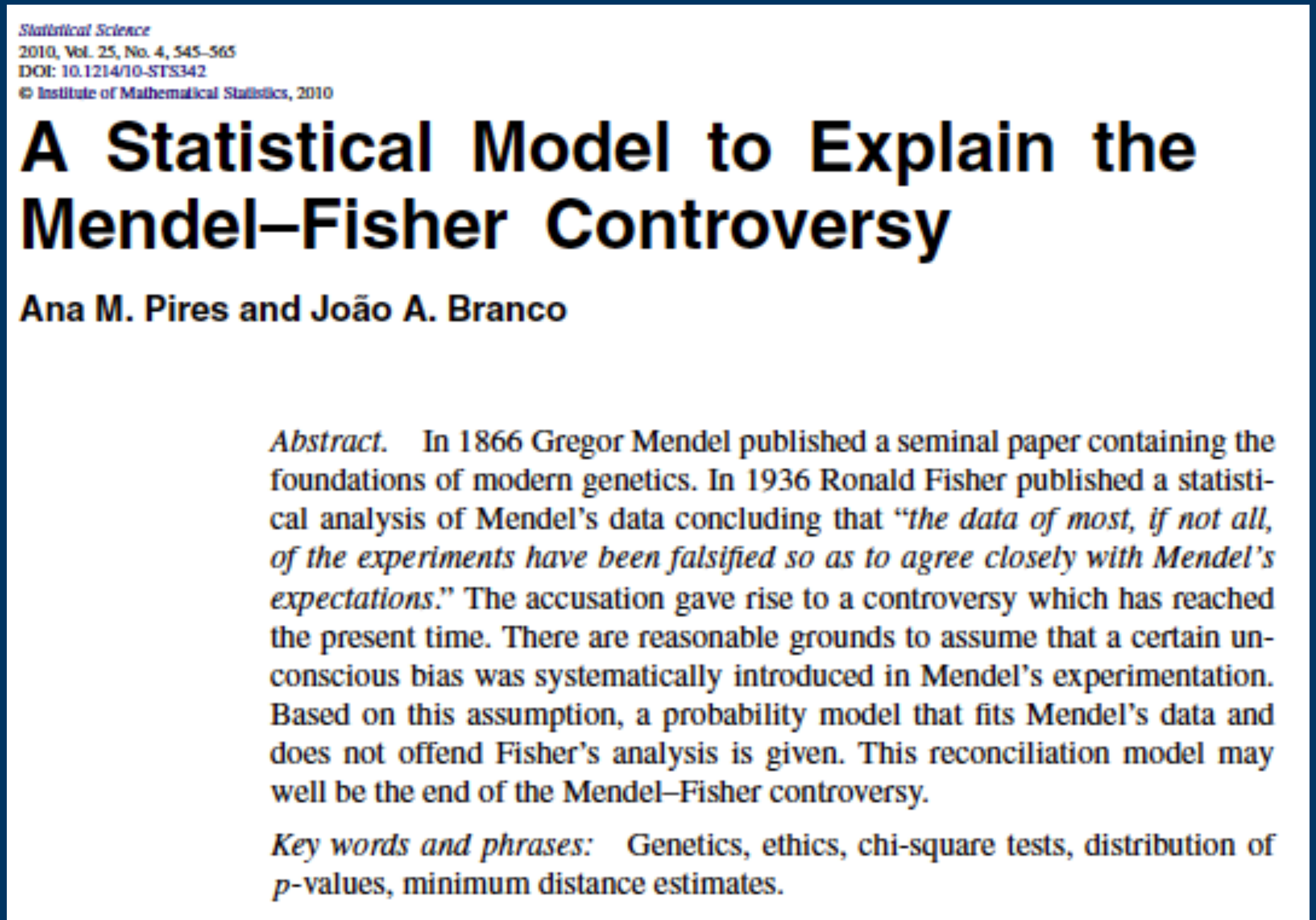
* For further commentary on Mendel's work written by Fisher in 1955, see Experiments in Plant Hybridisation: Gregor Mendel. (Ed. J.H. Bennett) Edinburgh: Oliver & Boyd, 1965. As indicated there, all of the years given in Fisher's (1936) reconstruction of the timing of Mendel's experimental programme must be reduced by one.

Annals of Science, 1: 115-137, (1936).



detail by his paper as a whole. Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected. This possibility is supported by independent evidence that the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations.

Mendel-Fisher Controversy



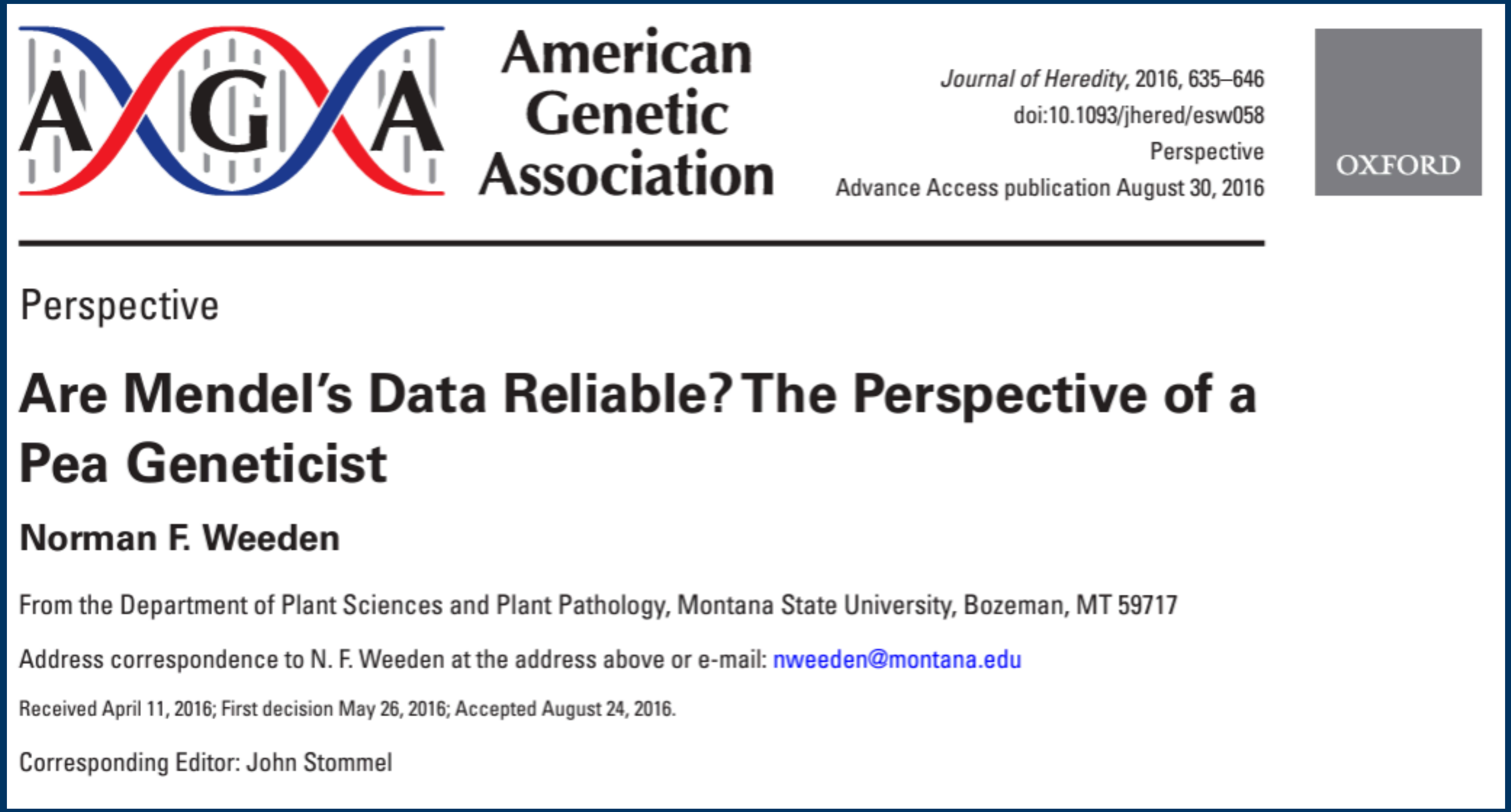
Statistical Science
2010, Vol. 25, No. 4, 545–565
DOI: 10.1214/10-STS342

A Statistical Model to Explain the Mendel–Fisher Controversy

Ana M. Pires and João A. Branco

Abstract. In 1866 Gregor Mendel published a seminal paper containing the foundations of modern genetics. In 1936 Ronald Fisher published a statistical analysis of Mendel's data concluding that "*the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations*." The accusation gave rise to a controversy which has reached the present time. There are reasonable grounds to assume that a certain unconscious bias was systematically introduced in Mendel's experimentation. Based on this assumption, a probability model that fits Mendel's data and does not offend Fisher's analysis is given. This reconciliation model may well be the end of the Mendel–Fisher controversy.

Key words and phrases: Genetics, ethics, chi-square tests, distribution of p-values, minimum distance estimates.



American Genetic Association

Journal of Heredity, 2016, 635–646
doi:10.1093/jhered/esw058
Perspective
Advance Access publication August 30, 2016

OXFORD

Perspective

Are Mendel's Data Reliable? The Perspective of a Pea Geneticist

Norman F. Weeden

From the Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT 59717

Address correspondence to N. F. Weeden at the address above or e-mail: nweeden@montana.edu

Received April 11, 2016; First decision May 26, 2016; Accepted August 24, 2016.

Corresponding Editor: John Stommel

Exercise 3:

$$X \rightsquigarrow F \Rightarrow \begin{cases} Y = F(X) \rightsquigarrow \text{Uniform}(0,1) \\ Y = 1 - F(X) \rightsquigarrow \text{Uniform}(0,1) \end{cases}$$

Create a pooled sample of the p-values from exercises 1 and 2 and test whether the p-values are coming from an Uniform distribution.

Draw your conclusions.

First creation of Genotype-Mapping

Genotype	Phenotype
AA	A
Aa	A
aa	a

If you know the genotype, then you know the phenotype

One gene that controls a single binary trait

Mendel triggered the scientific curiosity

What is actually a gene and an allele?

What is the gene involved?

Is it possible to derive genotype-phenotype rules for other type of traits such as the occurrence of a given disease or height?

Some useful concepts

Gene = a stretch of DNA located in a chromosome. The stretches encodes a protein

Allele = variant in the DNA sequence of a gene

Chromosome = a long DNA molecule that contains genetic information of an organism

Genome = the set of all the chromosomes that enables the creation of life

Human Genome = 1-23 autosomal chromosomes, X and Y sexual chromosomes

Beyond Mendelian

Genotype	→	Phenotype	Probabilities
AA		A or a	π_{AA}
Aa		A or a	π_{Aa}
aa		A or a	π_{aa}

Complete penetrance versus incomplete penetrance

What is the gene responsible for this trait?

Genetic mapping: general principle

What is the gene responsible for this trait?

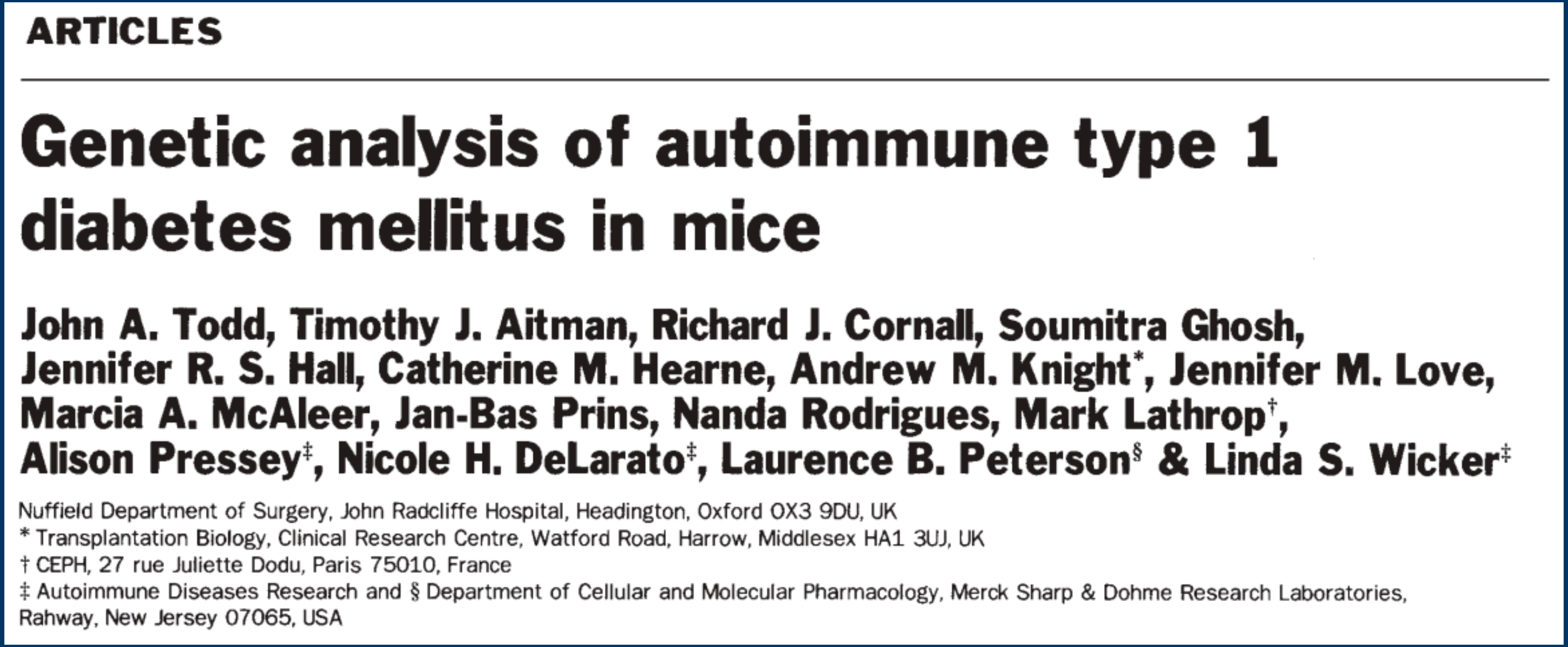
Use experimental cross a la Mendel.

Use genetic markers (with alleles a and b) at known location in the genome.

Test for association between the genotypes and the trait.

Example: Genetic mapping of Type 1 diabetes in mice

data_todd_1991.csv



ARTICLES

Genetic analysis of autoimmune type 1 diabetes mellitus in mice

John A. Todd, Timothy J. Aitman, Richard J. Cornall, Soumitra Ghosh, Jennifer R. S. Hall, Catherine M. Hearne, Andrew M. Knight[*], Jennifer M. Love, Marcia A. McAleer, Jan-Bas Prins, Nanda Rodrigues, Mark Lathrop[†], Alison Pressey[‡], Nicole H. DeLarato[‡], Laurence B. Peterson[§] & Linda S. Wicker[‡]

Nuffield Department of Surgery, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK
* Transplantation Biology, Clinical Research Centre, Watford Road, Harrow, Middlesex HA1 3UJ, UK
† CEPH, 27 rue Juliette Dodu, Paris 75010, France
‡ Autoimmune Diseases Research and § Department of Cellular and Molecular Pharmacology, Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065, USA

Genetic mapping: type 1 diabetes in mice

High incidence | Low incidence

NOD x (B10.H-2g x NOD) F1

NN | NB

Progeny

NN | NB

Each genetic marker 50% | 50%

53 genetic markers across the genome

Exercise 4:

Use an appropriate statistical test and test whether second Mendel’s law apply to the data.

Which the genetic marker has the highest association with the trait?

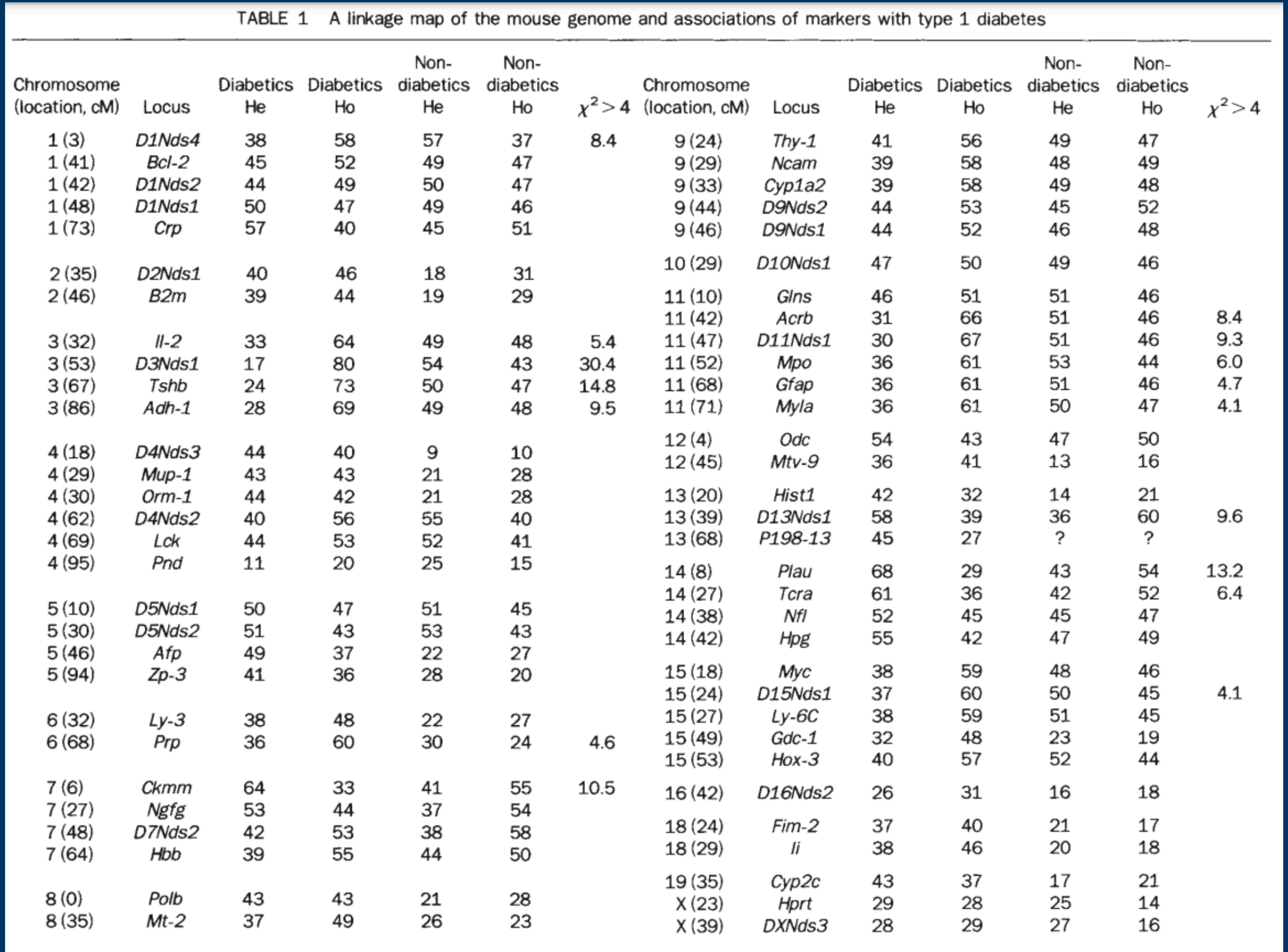


TABLE 1 A linkage map of the mouse genome and associations of markers with type 1 diabetes

Chromosome (location, cM)	Locus	Diabetics He	Diabetics Ho	Non-diabetics He	Non-diabetics Ho	$\chi^2 > 4$
1 (3)	*D1Nds4*	38	58	57	37	8.4
1 (41)	*Bcl-2*	45	52	49	47	
1 (42)	*D1Nds2*	44	49	50	47	
1 (48)	*D1Nds1*	50	47	49	46	
1 (73)	*Crp*	57	40	45	51	
2 (35)	*D2Nds1*	40	46	18	31	
2 (46)	*B2m*	39	44	19	29	
3 (32)	*Il-2*	33	64	49	48	5.4
3 (53)	*D3Nds1*	17	80	54	43	30.4
3 (67)	*Tshb*	24	73	50	47	14.8
3 (86)	*Adh-1*	28	69	49	48	9.5
4 (18)	*D4Nds3*	44	40	9	10	
4 (29)	*Mup-1*	43	43	21	28	
4 (30)	*Orm-1*	44	42	21	28	
4 (62)	*D4Nds2*	40	56	55	40	
4 (69)	*Lck*	44	53	52	41	
4 (95)	*Pnd*	11	20	25	15	
5 (10)	*D5Nds1*	50	47	51	45	
5 (30)	*D5Nds2*	51	43	53	43	
5 (46)	*Afp*	49	37	22	27	
5 (94)	*Zp-3*	41	36	28	20	
6 (32)	*Ly-3*	38	48	22	27	
6 (68)	*Prp*	36	60	30	24	4.6
7 (6)	*Ckmm*	64	33	41	55	10.5
7 (27)	*Ngfg*	53	44	37	54	
7 (48)	*D7Nds2*	42	53	38	58	
7 (64)	*Hbb*	39	55	44	50	
8 (0)	*Polb*	43	43	21	28	
8 (35)	*Mt-2*	37	49	26	23	
9 (24)	*Thy-1*	41	56	49	47	
9 (29)	*Ncam*	39	58	48	49	
9 (33)	*Cyp1a2*	39	58	49	48	
9 (44)	*D9Nds2*	44	53	45	52	
9 (46)	*D9Nds1*	44	52	46	48	
10 (29)	*D10Nds1*	47	50	49	46	
11 (10)	*Glns*	46	51	51	46	
11 (42)	*Acrb*	31	66	51	46	8.4
11 (47)	*D11Nds1*	30	67	51	46	9.3
11 (52)	*Mpo*	36	61	53	44	6.0
11 (68)	*Gfap*	36	61	51	46	4.7
11 (71)	*Myla*	36	61	50	47	4.1
12 (4)	*Odc*	54	43	47	50	
12 (45)	*Mtv-9*	36	41	13	16	
13 (20)	*Hist1*	42	32	14	21	
13 (39)	*D13Nds1*	58	39	36	60	9.6
13 (68)	*P198-13*	45	27	?	?	
14 (8)	*Plau*	68	29	43	54	13.2
14 (27)	*Tcra*	61	36	42	52	6.4
14 (38)	*Nfl*	52	45	45	47	
14 (42)	*Hpg*	55	42	47	49	
15 (18)	*Myc*	38	59	48	46	
15 (24)	*D15Nds1*	37	60	50	45	4.1
15 (27)	*Ly-6C*	38	59	51	45	
15 (49)	*Gdc-1*	32	48	23	19	
15 (53)	*Hox-3*	40	57	52	44	
16 (42)	*D16Nds2*	26	31	16	18	
18 (24)	*Fim-2*	37	40	21	17	
18 (29)	*Ii*	38	46	20	18	
19 (35)	*Cyp2c*	43	37	17	21	
X (23)	*Hprt*	29	28	25	14	
X (39)	*DXNds3*	28	29	27	16	

Discussion:

What is the statistical challenge of genetic mapping?

Multiple testing problem

In absence of association

$\alpha = 0.05$ $m =$ number of genetic markers (statistical tests)

$Y =$ Number of significant tests $| H_0, \alpha = 0.05 \rightsquigarrow$ Binomial$(m, p = \alpha)$

Expected number of false positive associations

$$E[Y | H_0, \alpha] = m \times \alpha$$

$$E[Y | H_0, \alpha] = 53 \times 0.05 = 2.65$$

Dealing with multiple testing (classical methods)

Redefine the type I error for the overall analysis

$$P[Y \geq 1 \,|\, H_0, \alpha^*] = \alpha$$

$$1 - (\alpha^*)^m = \alpha \Leftrightarrow \alpha^* = 1 - (1 - \alpha)^{1/m}$$

Sidak-Dunn correction

$$\alpha^* = 1 - (1 - \alpha)^{1/53} \approx 0.00084$$

$$E[Y \,|\, H_0, \alpha^*] = \alpha$$

$$m \times \alpha^* = \alpha \Leftrightarrow \alpha^* = \frac{\alpha}{p}$$

Bonferroni correction

$$\alpha^* = \frac{0.05}{53} = 0.00082$$

In the previous exercise, was the strongest association statistically significant controlling for multiple testing?

Mendelian Genetics

Single gene, single binary trait

Complete penetrance

Rules of dominance/recessiveness

Non-Mendelian Genetics

Complex binary traits

Presence of common diseases (diabetes, multiple sclerosis, COVID-19 lethality)

Categorical traits

Eye color

Multiple interacting gene involved

Quantitative traits

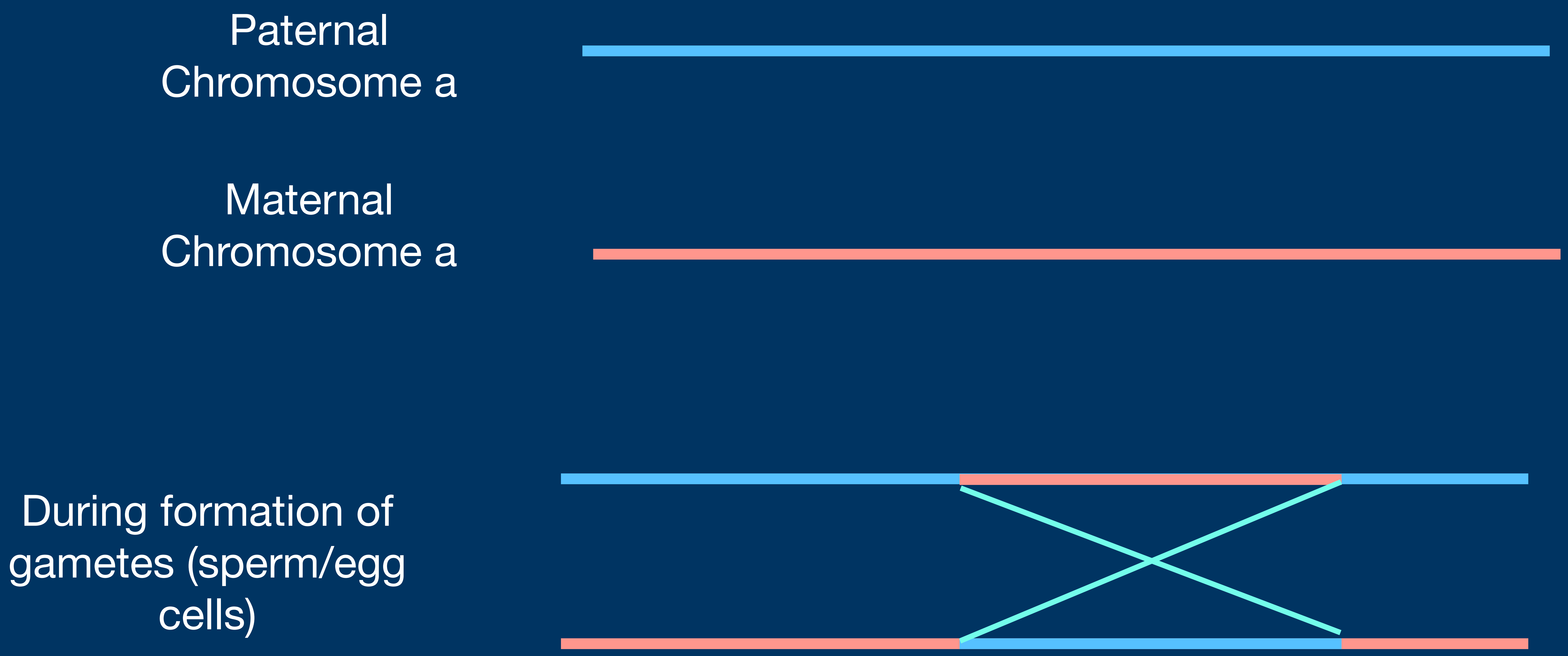
Height, Haemogloblin levels, blood pressure

Basic question

What are the genes involved and what is their action on the phenotype?

The identification of the genes involved is expected to improve human health by targeting the causative genes

Recombination



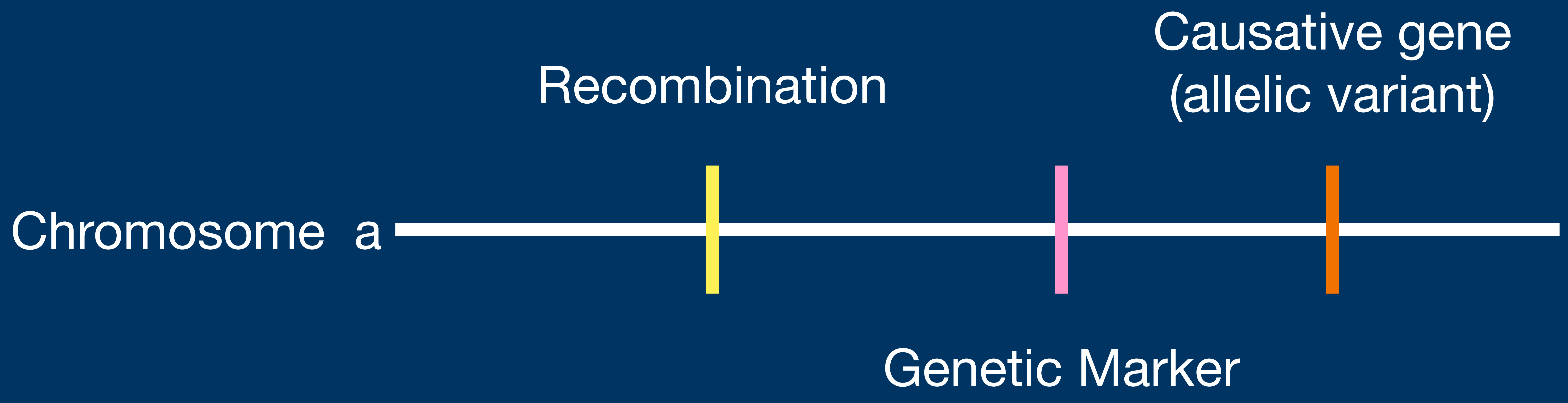
Paternal Chromosome a

Maternal Chromosome a

During formation of gametes (sperm/egg cells)

Linkage and recombination

Complete linkage

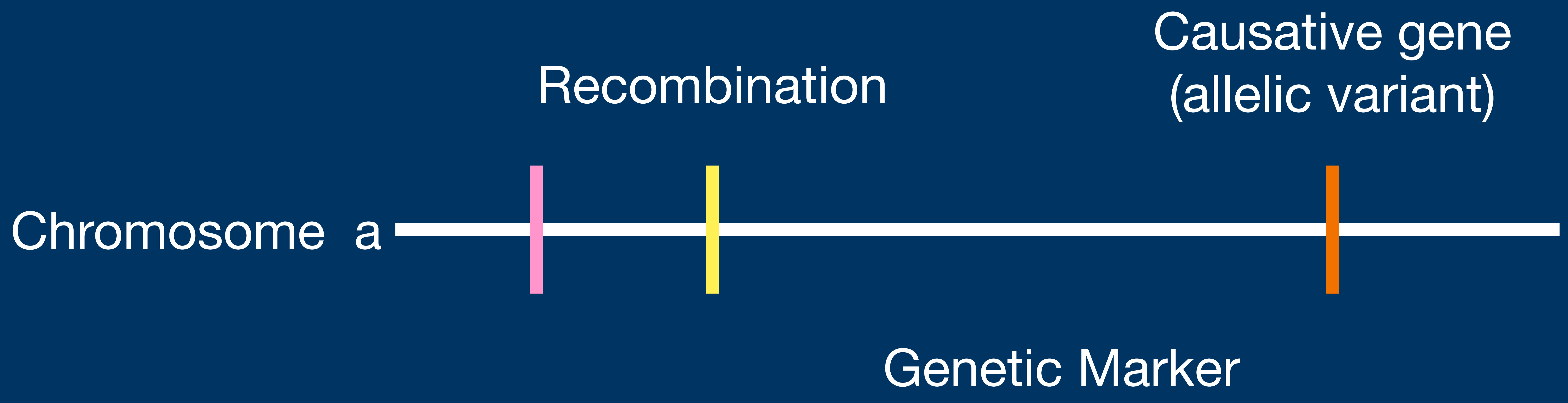


If a genetic marker and the causative gene are physically close to each other, then the genetic marker and the gene are inherited together.

The genetic marker fully represents the true statistical association between the causative gene and the phenotype.

Linkage and recombination

Incomplete linkage

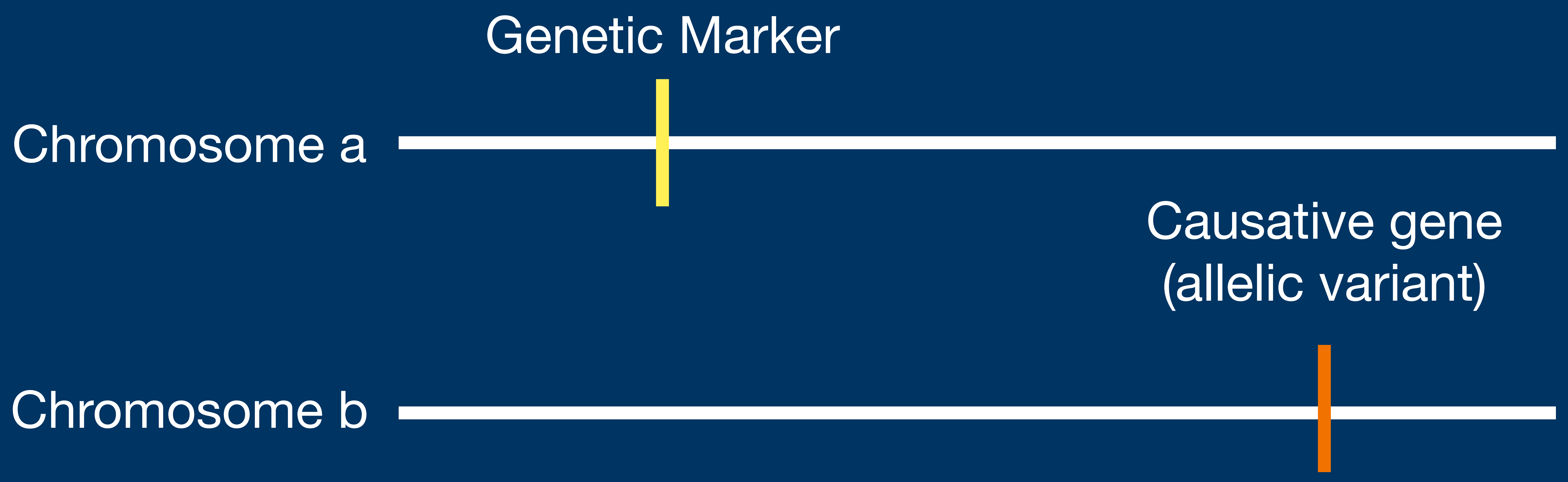


If a genetic marker and the causative gene are not physically close to each other, then a recombination might occur during gametogenesis. This recombination is passed onto the offsprings

The genetic marker partially represents the true statistical association between the causative gene and the phenotype.

Linkage and recombination

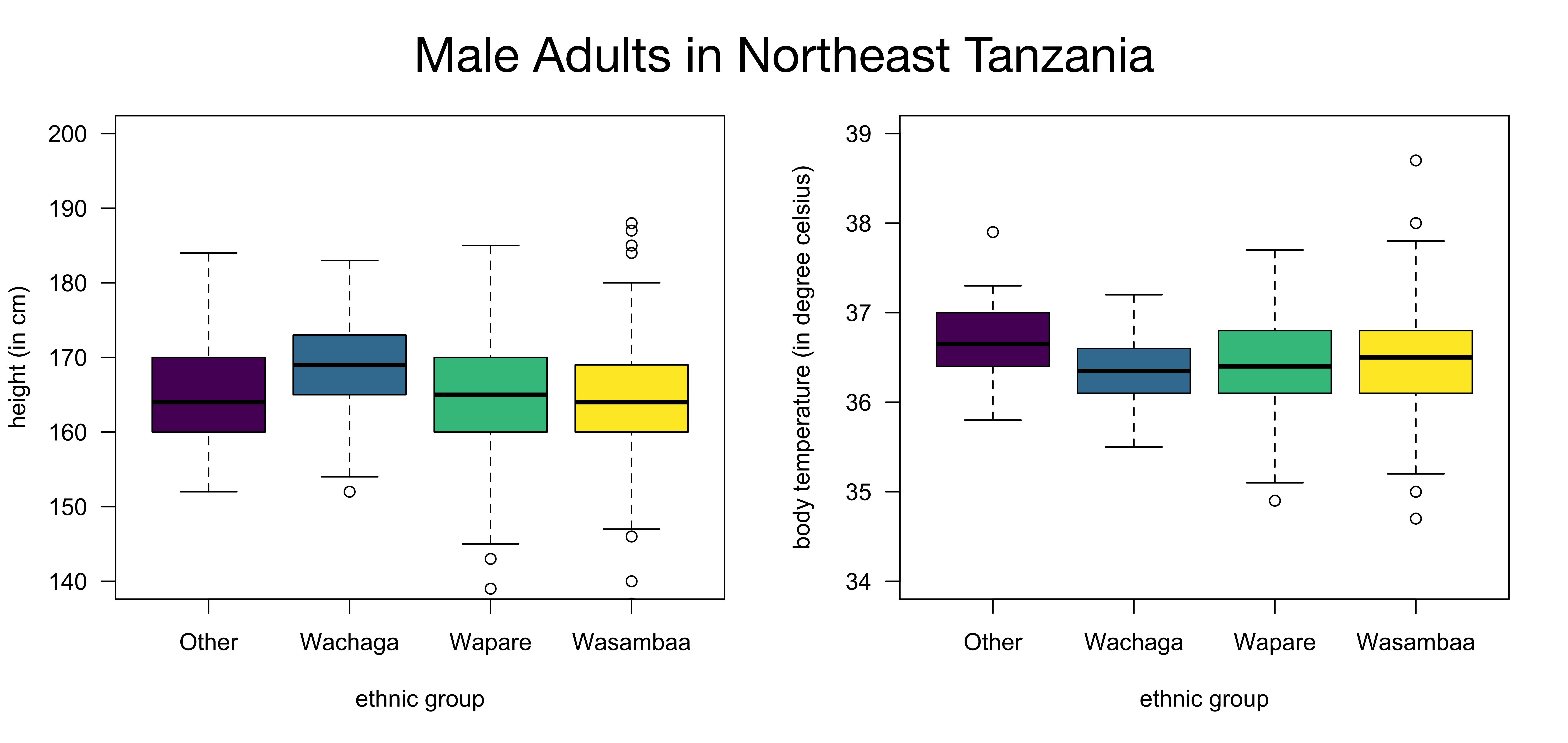
No linkage



If a genetic marker and the causative gene are in different chromosomes, there is no association between the genetic marker and the causative gene due to independent segregation of chromosomes.

The genetic marker is not associated with the phenotype.

A bit about quantitative traits



What are the genes controlling the height and body temperature of these individuals?

Fisher’s infinitesimal (or polygenic) model

A quantitative trait is affected by a large number of alleles located at different genes. The effect of these alleles is additive on the quantitative.

$$Y_i = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j X_i \rightsquigarrow ?$$

where

X_{ij} is the number of alleles in genotype at gene j in individual i

α_0 is the overall average of environmental factors and alleles located at other genes

α_j is the phenotype effect ofadding an allele to the genotype at gene j

Practical implications of Fisher’s infinite allele model

The genotype of an individual is converted in the number of a given allele (no rules of dominance and recessiveness as proposed by Mendel)

Interaction among different causative genes is discarded

↓

We can simplify the analysis of multiple genetic markers by analysis each genetic marker separately

↓

Additive model for the analysis of single genetic marker

Additive model for a single genetic marker for diploid organisms (humans!)

Y_i = random variable for the quantitative trait in individual i

$$Y_i \,|\, \mu_i, \sigma \rightsquigarrow N(\mu_i, \sigma^2)$$

X_i = the number of a given allele in the genotype of individual i for the genetic marker

$$X_i^* \in \{'aa', 'aA', 'AA'\} \longrightarrow X_i \in \{0,1,2\}$$

$\mu_i = \alpha + \alpha_1 X_i$

Assume sampling of unrelated individuals from the population

Additive model is a simple linear regression using a covariate with three numeric levels

Note: if sampling includes individuals from the same family, we need to include a random effect to contemplate the correlation among individuals due to genetic relatedness (similar to repeated measurement models - linear mixed models)

Testing the effect of a marker on the phenotype

$H_0 : \alpha_1 = 0$ versus $H_1 : \alpha_1 \neq 0$

Wald's Score test

$$S = \frac{\hat{\alpha}_1}{se(\hat{\alpha}_1)} \,|\, H_0 \rightsquigarrow Normal(\mu = 0, \sigma^2 = 1)$$

Wilks' likelihood ratio test

$$\Lambda = (-2)\frac{L(\hat{\alpha}_0^*)}{L(\hat{\alpha}_0, \hat{\alpha}_1)} \,|\, H_0 \rightsquigarrow \chi^2_{(1)}$$

$L(\hat{\alpha}_0^*) =$ maximised log-likelihood of the regression model without the covariate

$L(\hat{\alpha}_0, \hat{\alpha}_1) =$ maximised log-likelihood of the regression model with the covariate

Extending the additive model

$$Y_i \mid \mu_i, \sigma \rightsquigarrow N(\mu_i, \sigma^2)$$

Under the assumption of sampling unrelated individuals

$$\mu_i = \alpha + \alpha_1 X_i + \underbrace{\beta_1 Z_{1i} + \cdots + \beta_p Z_{pi}}_{\text{Non genetic covariates}}$$

Under the assumption of sampling unrelated individuals

$$\mu_i = \alpha + \underbrace{\alpha_1 X_{1i} + \cdots + \alpha_1 X_{mi}}_{\text{Associated genetic markers}} + \underbrace{\beta_1 Z_{1i} + \cdots + \beta_p X_{pi}}_{\text{Non-genetic covariates}}$$