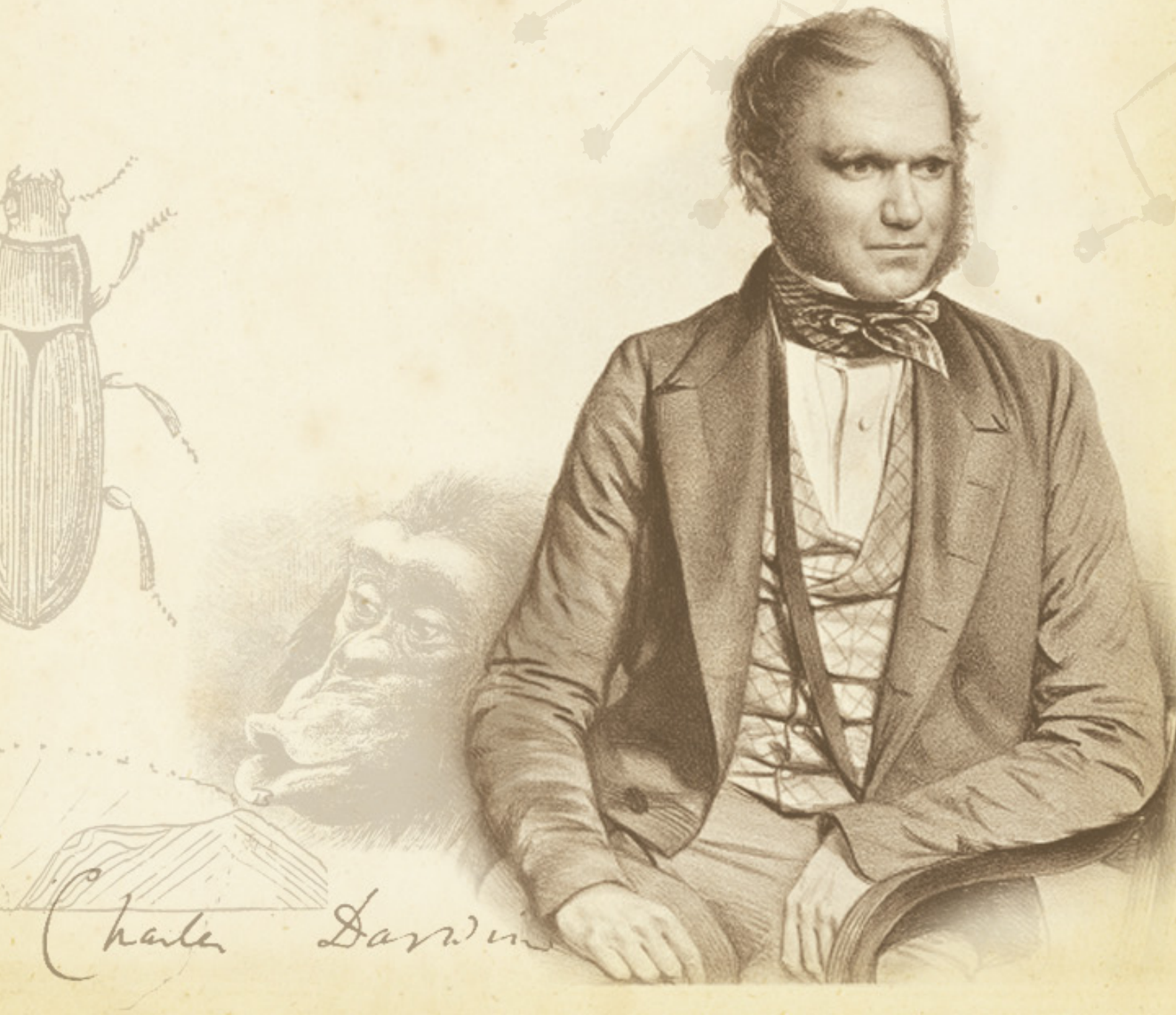
TEACHER'S GUIDE

DNA to Darwin Case Study

Lactose tolerance

Version 1.1

Steve Cross, Bronwyn Terrill and colleagues
Wellcome Trust Sanger Institute
Hinxton



Lactose tolerance

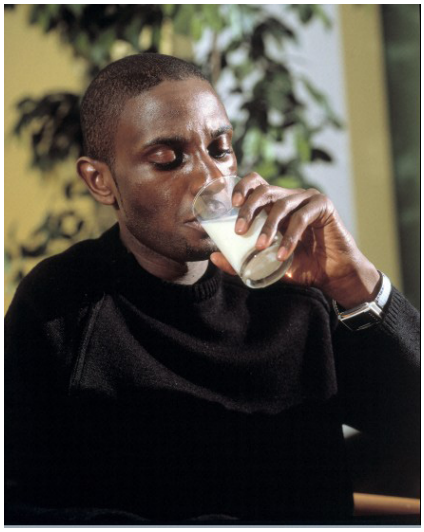


IMAGE FROM: Wellcome Images.

The majority of people living today are unable to digest lactose in adulthood. In this activity students will analyse DNA data from different human populations to work out which variations in the human genome have enabled some adult humans to consume dairy products without adverse side-effects. Students will also discover which genetic variants are associated with lactose tolerance (or lactase persistence) in different populations. A statistical test (Chi-squared) will be used to work out whether specific genetic changes have significant effects on lactase persistence.

Lactose, the sugar found in dairy products and milk, is metabolised by the enzyme lactase (full name: lactase-phlorizin hydrolase or LPH). Lactase splits apart the two sugars (galactose and glucose) that make up lactose. The gene for lactase is known as *LCT*, and it is mostly expressed in cells in the epithelium of the small intestine. Although nearly everyone can produce the enzyme at birth, the majority of people outside Northern Europe lose the ability to produce a lot of it at some point during the first years of life. This can be referred to as lactase nonpersistance, adult hypolactasia or, more commonly, lactose intolerance.

Lactase persistence is thought to be an example of recent positive selection for an evolutionary change in people from milk-drinking cultures. Research in northern European populations with a history of animal husbandry has found variants in a control (promoter) region about 14 kb upstream from the *LCT* gene itself. Finnish studies identified two potential variants, or single nucleotide polymorphisms (SNPs); one of them, at position –13910 was 100% correlated with lactase persistence. A 'C' at that position is now known to be the ancestral, lactase non-persistent form, while a 'T' is the lactose tolerant, milk-drinking form (Enattah, 2002). This muation is thought to have occurred between 8,000 and 9,000 years ago.

More recent studies have identified three additional mutations that allow adult milk-drinking in East African populations (Tishkoff *et al*, 2007). These are a G/C variant at –14010, T/G variant at –13915 and C/G at –13907. The variant at –14010 is statistically significant, and it has been shown through biochemical studies to alter lactase activity in modified cells. In this case, a 'G' at –14010 is the milk-drinking form. This variant has been dated to between 2,700 and 6,800 years ago, making it newer than the European variant.

No African populations have been found with the European variant, meaning that populations from outside Africa that are lactose-tolerant have one form, whereas lactose-tolerant Africans can have three other forms. These variants appear to have evolved independently — the newer mutations appearing in Africa — and stabilised as the populations took more milk into their diets. Possible selective factors are the nutritional benefits of milk, and perhaps protection against disease related to low dietary calcium intake.

Scientific publications

Enattah, N.S. *et al* (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, **30**, 233–237. doi: 10.1038/ng826

Johns Hopkins University (2008) Lactase; LCT. *Online Mendelian Inheritance in Man*: **www.ncbi.nlm.nih.gov/omim/603202**

Kuokkanen, M *et al* (2006) Mutations in the translated region of the lactase gene (LCT) underlie congenital lactase deficiency. *American Journal of Human Genetics*, **78** (2) 339–344.

Swallow, D.M. (2003) Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics*, **37** (n) 197–219. doi: 10.1146/annurev.genet.37.110801.143820

Tishkoff, S.A. *et al* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, **39** (1) 31–40. doi:10.1038/ng1946

Weiss, K. (2004) The unkindest cup. *The Lancet*, **363** (9420) 1489–1490. doi:10.1016/S0140-6736(04)16195-3

Requirements

Software

The data is presented for students to analyse as a *Microsoft Excel* spreadsheet. *Excel* can also be used to calculate the expected and Chi-squared values, or students can be asked to work through the process manually using the tables in the spreadsheet provided.

Microsoft Excel

Students will need a computer equipped with *Microsoft Excel*.

Students' worksheets

Students will require copies of Student's Guide, pages 2–11.

Educational aims

The subject of lactase persistence is a fertile one for the classroom. This is partly due to people's personal experiences or knowledge of friends who are lactose intolerant. It may also be due to the relative simplicity of the concept that a variant that increases lactase activity could benefit an adult in a community that raises livestock.

However, other concepts could be explored here, such as:

- How different mutations can arise independently in different, isolated populations;
- evidence can be misconstrued based on the researcher's context *e.g.*, the original 'adult lactose intolerance is an abnormality' view because the research was centred in Europe.

Prerequisite knowledge

Students will need to understand the basic structure of DNA (the *Introductory Activities*, which are in a separate document, will be useful here). It is essential that they understand the Chi-squared test and are competent at using spreadsheets in *Microsoft Excel*.

Results from Exercise 1

Data for this exercise comes directly from Enattah *et al* (2002).

The **LCT_teacher_spreadsheet**, Exercise 1 worksheet also contains the results for the exercise shown below.

C/T –13910

Condition *vs* genotypes	CC	CT	TT	Totals
Lactase deficiency / lactose intolerant	59	0	0	59
Lactase persistence	0	63	74	137
Total samples	59	63	74	196

G/A –22018

Condition *vs* genotypes	GG	GA	AA	Totals
Lactase deficiency / lactose intolerant	53	6	0	59
Lactase persistence	0	63	74	137
Total samples	53	69	74	196

Enattah *et al* (2002) found that the C/T variant at position –13910 is 100% associated with lactase persistence. The G/A –22108 variant is 97% associated. Since this study, position –13910 has been shown to be 80–90% associated with lactase persistence throughout other European populations.

Answers to the questions on the worksheet

Page 6

a. C/T–13910
b. CT, TT (at position –13910)
c. CC (at position –13910)
d. 6 ÷ 196 × 100 = 3.06 %

Results from Exercise 2

Data for this exercise comes directly from Tishkoff *et al* (2007).

The **LCT_teacher_spreadsheet**, Exercise 2 worksheet calculates the Chi-squared values for each of the three populations and two positions; the student worksheets are filled in for you on the following three pages.

Answers to the questions on the worksheet

Page 11

e. **Chi-squared values**
Kenyans: 21.41 Tanzanians: 21.43 Afro-Asiatic Kenyans: 6.14

f. **Probabilities**
Kenyans: <0.005 Tanzanians: <0.005 Afro-Asiatic Kenyans: 0.25–0.1

g. The Kanyan and Tanzanian results are statistically significant; the Afro-Asiatic Kenyan result is not.

h. Yes: Kenyans and Tanzanians at position –14010.

i. Different statistically significant variants account for lactase persistence in different populations (–13910 in Northern Europeans, –14010 in East Africans). This means that the trait may have evolved independently (and more than once) in isolated populations.

j. Examples to be discussed with students could include:
- More samples from different populations with lactase persistence.
- Larger samples to try to identify more variants with small effects.
- Model organism or cellular expression of lactase using genetic modification to confirm the effects of these variants.
- Level of variation around the mutation (less variation could mean recent positive selection).

Population name: Kenyan Position of allele: –14010

Category	Observed (O)	Expected* (E) (see below)	O–E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Genotype 1 (CC) *LCT* persistent	12	8.72	3.28	10.76	1.23
Genotype 2 (CG) *LCT* persistent	48	36.06	11.94	142.56	3.95
Genotype 3 (GG) *LCT* persistent	38	53.12	–15.12	228.61	4.30
Genotype 1 (CC) *LCT* non-persistent	2	4.63	–2.63	6.92	1.49
Genotype 2 (CG) *LCT* non-persistent	10	19.14	–9.14	83.54	4.36
Genotype 3 (GG) *LCT* non-persistent	40	28.18	11.82	139.71	4.96
Genotype 1 (CC) Intermediate	3	3.56	–0.56	0.31	0.09
Genotype 2 (CG) Intermediate	12	14.72	–2.72	7.40	0.50
Genotype 3 (GG) Intermediate	25	21.68	3.32	11.02	0.51
TOTALS	190				21.41

* To calculate the Expected values, you will need to count the number of number of samples with a single genotype found across the entire population. This figure, divided by the total sample size will give you the prevalence of the genotype. Multiplying the prevalence of each form of the allele by the number of samples in each condition group, will give you the Expected values.

Category	Size of population (ignoring alleles)	Prevalence of allele	Expected
Genotype 1 (CC) *LCT* persistent	(12 + 2 + 3 = 17) ÷ 190	0.089 × (12 + 48 + 38 = 98)	8.72
Genotype 2 (CG) *LCT* persistent	(48 + 10 + 12 = 70) ÷ 190	0.368 × 98	36.06
Genotype 3 (GG) *LCT* persistent	(38 + 40 + 25 = 103) ÷ 190	0.542 × 98	53.12
Genotype 1 (CC) *LCT* non-persistent	17 ÷ 190	0.089 × (2 + 10 + 40 = 52)	4.63
Genotype 2 (CG) *LCT* non-persistent	70 ÷ 190	0.368 × 52	19.14
Genotype 3 (GG) *LCT* non-persistent	103 ÷ 190	0.542 × 52	28.18
Genotype 1 (CC) Intermediate	17 ÷ 190	0.089 × (3 + 12 + 25 = 40)	3.56
Genotype 2 (CG) Intermediate	70 ÷ 190	0.368 × 40	14.72
Genotype 3 (GG) Intermediate	103 ÷ 190	0.542 × 40	21.68

Population name: Tanzanian Position of allele: –14010

Category	Observed (O)	Expected* (E) (see below)	O–E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Genotype 1 (CC) *LCT* persistent	17	11.35	5.65	31.92	2.81
Genotype 2 (CG) *LCT* persistent	44	34.82	9.18	84.27	2.42
Genotype 3 (GG) *LCT* persistent	36	50.83	–14.83	219.93	4.33
Genotype 1 (CC) *LCT* non-persistent	5	9.36	–4.36	19.01	2.03
Genotype 2 (CG) *LCT* non-persistent	18	28.72	–10.72	114.92	4.00
Genotype 3 (GG) *LCT* non-persistent	57	41.92	15.08	227.41	5.42
Genotype 1 (CC) Intermediate	5	6.32	–1.32	1.74	0.28
Genotype 2 (CG) Intermediate	21	19.39	1.61	2.59	0.13
Genotype 3 (GG) Intermediate	28	28.30	–0.3	0.09	0.00
TOTALS	231				21.43

* To calculate the Expected values, you will need to count the number of number of samples with a single genotype found across the entire population. This figure, divided by the total sample size will give you the prevalence of the genotype. Multiplying the prevalence of each form of the allele by the number of samples in each condition group, will give you the Expected values.

Category	Size of population (ignoring alleles)	Prevalence of allele	Expected
Genotype 1 (CC) *LCT* persistent	(17 + 5 + 5 = 27) ÷ 231	0.117 × (17 + 44 + 36 = 97)	11.35
Genotype 2 (CG) *LCT* persistent	(44 + 18 + 21 = 83) ÷ 231	0.359 × 97	34.82
Genotype 3 (GG) *LCT* persistent	(36 + 57 + 28 = 121) ÷ 231	0.542 × 97	50.83
Genotype 1 (CC) *LCT* non-persistent	27 ÷ 231	0.117 × (5 + 18 + 57 = 80)	9.36
Genotype 2 (CG) *LCT* non-persistent	83 ÷ 231	0.359 × 80	28.72
Genotype 3 (GG) *LCT* non-persistent	121 ÷ 231	0.542 × 80	41.92
Genotype 1 (CC) Intermediate	27 ÷ 231	0.117 × (5 + 21 + 28 = 54)	6.32
Genotype 2 (CG) Intermediate	83 ÷ 231	0.359 × 54	19.39
Genotype 3 (GG) Intermediate	121 ÷ 231	0.542 × 54	28.30

Population name: Afro-Asiatic Kenyans Position of allele: –13915

Category	Observed (O)	Expected* (E) (see below)	O–E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Genotype 1 (GG) *LCT* persistent	2	1.12	0.88	0.77	0.69
Genotype 2 (GT) *LCT* persistent	7	4.45	2.55	6.50	1.46
Genotype 3 (TT) *LCT* persistent	25	28.42	–3.42	11.70	0.41
Genotype 1 (GG) *LCT* non-persistent	0	0.43	–0.43	0.18	0.43
Genotype 2 (GT) *LCT* non-persistent	1	1.70	–0.7	0.49	0.29
Genotype 3 (TT) *LCT* non-persistent	12	10.87	1.13	1.28	0.12
Genotype 1 (GG) Intermediate	0	0.46	–0.46	0.21	0.46
Genotype 2 (GT) Intermediate	0	1.83	–1.83	3.35	1.83
Genotype 3 (TT) Intermediate	14	11.70	2.30	5.29	0.45
TOTALS	61				6.14

* To calculate the Expected values, you will need to count the number of number of samples with a single genotype found across the entire population. This figure, divided by the total sample size will give you the prevalence of the genotype. Multiplying the prevalence of each form of the allele by the number of samples in each condition group, will give you the Expected values.

Category	Size of population (ignoring alleles)	Prevalence of allele	Expected
Genotype 1 (GG) *LCT* persistent	(2 + 0 + 0 = 2) ÷ 61	0.033 × (2 + 7 + 25 = 34)	1.12
Genotype 2 (GT) *LCT* persistent	(7 + 1 + 0 = 8) ÷ 61	0.131 × 34	4.45
Genotype 3 (TT) *LCT* persistent	(25 + 12 + 14 = 51) ÷ 61	0.836 × 34	28.42
Genotype 1 (GG) *LCT* non-persistent	2 ÷ 61	0.033 × (0 + 1 + 12 = 13)	0.43
Genotype 2 (GT) *LCT* non-persistent	8 ÷ 61	0.131 × 13	1.70
Genotype 3 (TT) *LCT* non-persistent	51 ÷ 61	0.836 × 13	10.87
Genotype 1 (GG) Intermediate	2 ÷ 61	0.033 × (0 + 0 + 14 = 14)	0.46
Genotype 2 (GT) Intermediate	8 ÷ 61	0.131 × 14	1.83
Genotype 3 (TT) Intermediate	51 ÷ 61	0.836 × 14	11.70