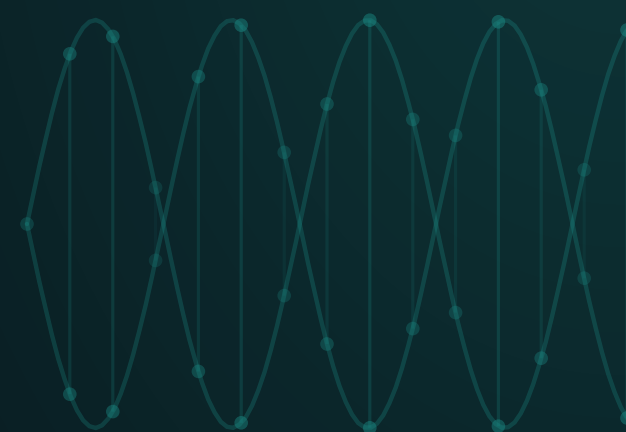


Lesson: genes and rsIDs, and what a catalogue actually measures



Biology

Genetics

Age 16+

Introductory

A ready-to-run genetics lesson for age 16 and up, free to use in any school, college or university. One book everybody owns a copy of: a chapter is a gene, and an rsID names a letter that differs between copies. Then students count those letters in a real catalogue and find the counts measure research attention rather than biology.

Free to read, use and adapt. No account, nothing to buy, nothing collected from anyone.

Full lesson, figures and updates:

<https://www.mygenelog.com/updates/lesson-genes-and-rsids-as-code>

What this is. A ready-to-run genetics lesson, free to use and adapt in any school, college or university. Nothing to buy from us, no account, no data collected from anyone.

- Level: works from about age 16 through undergraduate. No genetics needed. Some familiarity with the idea of a source file helps but is not required — the analogy is explained from scratch.
- Time: 45–60 minutes. The counting activity takes about fifteen.
- You need: a browser per student or per pair, and a board to write numbers on. No genome, no sample, no test of any kind.
- It covers: what a gene is and what an rsID is, how many of each there are, why a catalogue is not the same as reality, and how research attention leaves a fingerprint in data that looks purely biological.

The whole thing in five lines

Everyone alive owns a copy of the same book.

- The book is your DNA. Three billion letters, bound into 23 volumes.
- A chapter is a gene — the part that explains how to make one thing. There are about

20,000 chapters.

- Your copy and mine are almost identical. Every so often, one letter is different: mine reads C where yours reads T.
- Scientists gave every one of those differing spots a permanent number. That number is an rsID, like rs1815739.
- So an rsID is the name of a place, not the letter you have there. Which letter you carry is a separate question, and it is called your genotype.

That is the lesson's whole vocabulary. Everything below is either a closer look at one of those five lines, or the thing students actually do with them.

The one sentence to keep. A gene is a chapter; an rsID is the number given to a single letter inside it. One chapter holds many such letters — which is why "I have the BRCA1 gene" cannot be right on its own. Everybody has BRCA1. What differs is which letters are in their copy of it.

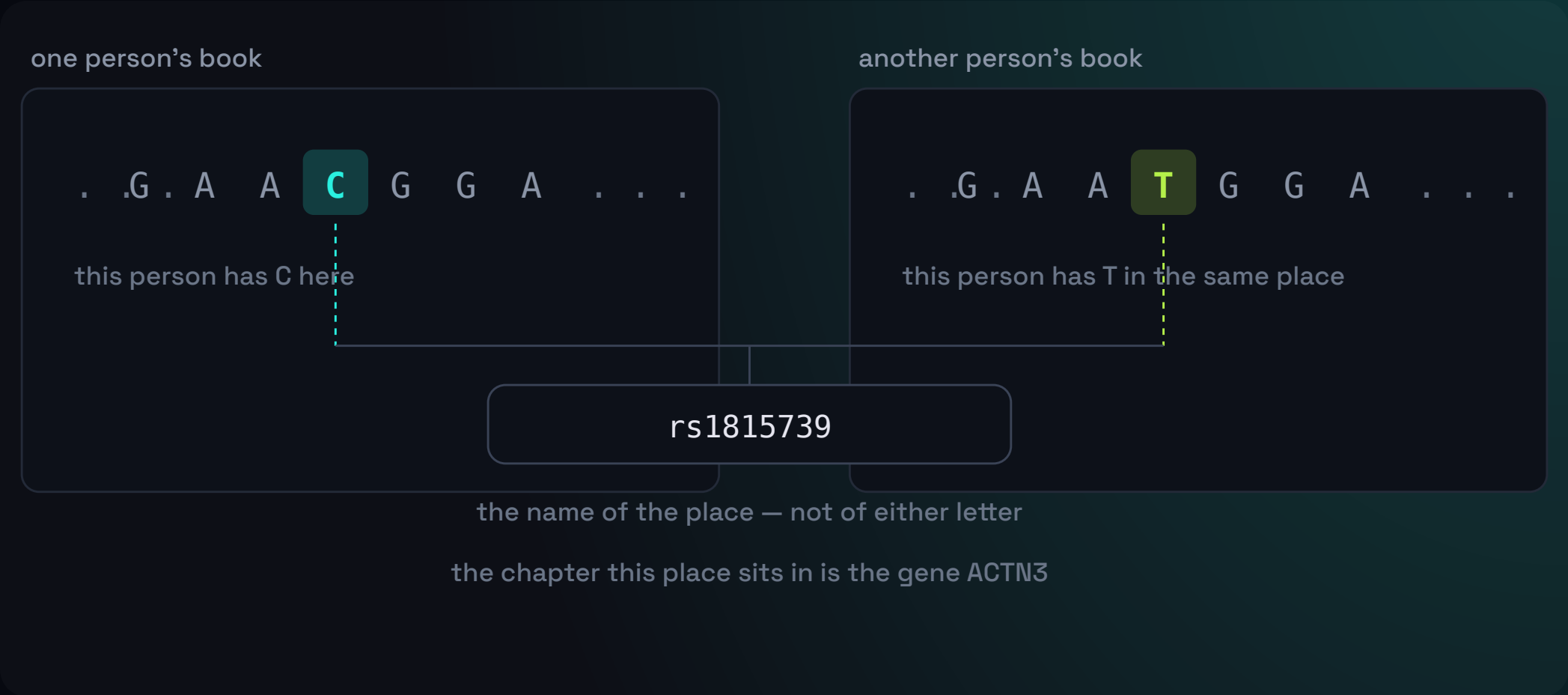
Two different "twos" — do not mix them up

This is the commonest confusion in the whole topic, and it is worth two minutes before the picture rather than an argument after it. The word "two" is about to mean two different things.

- Two people. Your book and mine are almost identical, and differ at the odd letter. That is what an rsID is for: it names a place where people differ.
- Two sets inside one person. You own two complete sets of the book — one inherited from your mother, one from your father. So when you open any single position you find not one letter but two. That pair is your genotype, and it is why a test result is written C/C, C/T or T/T rather than as a single letter.

The figure below is the first kind: two different people, one letter apart. Keep the second kind in mind, because it is the reason a result has two letters in it.

The same idea, drawn



Schematic, not to scale. Every letter is the same in both books except one. rs1815739 is the name of that place — it does not tell you which letter anybody has there. Which letter you carry is your genotype, and it is a separate question.

In the book	In a genome	What it is
The letters, and the paper they are printed on	DNA	Four letters — A, T, G, C — paired into a ladder and twisted. That twist is the famous double helix. It is the paper, not the writing.
The 23 volumes it is bound into	Chromosomes	Two metres of DNA wound tight enough to fit inside a cell you cannot see. You carry 23 pairs.
A chapter that explains how to make one thing	Gene	A stretch of the text carrying the instructions for one product, usually a protein. About 20,000 of them.
A letter that reads differently in one person's book than in another's	SNP	A single position where people differ. Almost every letter is identical in everyone; these are the rare ones that are not.
The catalogue number for that exact spot	rsID	A permanent name so that everybody, everywhere, means the same position. rs1815739 is the same place in every paper on Earth. It names the place, not the letter you happen to have there.

So the containment runs one way: a chapter contains many such positions, and a position sits inside at most one chapter — often none at all, because most of the book

is not chapters.

Where the analogy breaks, and it matters. A book is read front to back and says the same thing to every reader. This one is not. Different chapters are read in different cells — a muscle cell and a liver cell read the same book and use different parts of it. Most of a chapter’s text is cut out before the instruction is used. And one chapter can yield several different products depending on which parts are kept. Push the metaphor past “chapter and letter” and it will start telling you things that are not true.

If you have written code, there is a tighter version of this
A gene is a source file; an rsID names a line inside it that differs between forks. The same caution applies and applies harder: code is executed as written, top to bottom, and DNA is not.

How many of each are there?

The numbers are worth saying out loud, because they are not the numbers most people expect.

Thing	Roughly how many, in one human
Chromosomes	46, in 23 pairs
DNA letters	about 3 billion per copy
Length of DNA in a single cell, unwound	about 2 metres
Protein-coding genes	about 20,000
Genes that are transcribed but make no protein	tens of thousands more
rsIDs described so far, across all people	hundreds of millions

The 20,000 is the surprise. Before the genome was sequenced, serious estimates ran to 100,000 and higher, on the reasoning that a complex organism needs a lot of instructions. The count came back far lower, and the resolution is that the instructions are reused: one gene can yield several products, and a great deal of the genome is regulation deciding when the rest runs.

Do this: count the rsIDs in a gene

This is the part where the two words stop being definitions and start being something you can measure. Everything here uses the free pages on this site; nothing is

submitted, and no account is needed.

Before you start — the one rule. Nobody brings their own genetic data to this lesson, and nobody is asked whether they carry anything. There is no test here and no result about any person in the room. The activity is about a public catalogue, not about you. If a student asks about their own DNA, the honest answer is that a class is the wrong place to find out and a doctor or genetic counsellor is the right one.

1. Predict first, before opening anything

Write two numbers on the board, agreed by the class before any searching:

- How many rsIDs do you expect to find inside BRCA1, the gene most people in the room have heard of?
- How many inside PITX2, which most of them have not?

Almost every class predicts the famous gene will have more. Write the prediction down. The prediction being wrong is the lesson, and it only works if it is committed to first.

2. Count what this catalogue actually holds

Search each gene symbol on this site and count the variant pages returned. As of writing, the catalogue holds 2,976 variants across 2,076 genes, and the counts look like this:

Gene	Known for	rsIDs here	What those entries are about
PITX2	heart rhythm	15	atrial fibrillation — all fifteen
FTO	body weight	10	body mass and related measures
ABO	blood group	9	a range of traits
APOE	Alzheimer risk	4	cholesterol and Alzheimer disease
TAS2R38	bitter taste	3	tasting PTC
MTHFR	folate processing	2	folate metabolism
CFTR	cystic fibrosis	2	cystic fibrosis
ACTN3	muscle fibre type	1	muscle fibre type
TP53	the most-studied gene in cancer	1	basal cell carcinoma

Gene	Known for	rsIDs here	What those entries are about
BRCA1	breast and ovarian cancer risk	1	age at first period

Read the last column before going on. It is doing more work than the counts are.

3. The control: ask what the number is measuring

Put the prediction beside the result. TP53 is the most-studied gene in all of cancer biology and it has one entry here. PITX2 has fifteen. Now the question that makes this a science lesson rather than a browsing exercise: does PITX2 vary more between people than TP53 does?

It does not. The count measures how this catalogue was built. Almost everything here comes from genome-wide association studies, which scan the genome for positions that differ between people who have a trait and people who do not. Atrial fibrillation has been studied that way many times, PITX2 comes up every time, and each study leaves entries behind. All fifteen PITX2 rows are the same trait, found again and again.

The two bottom rows make the point sharper than any explanation could.

TP53 is arguably the most important gene in cancer research, and its single entry here is for basal cell carcinoma. That is not an error. TP53 matters mostly through mutations that arise in a tumour during a person’s life, which are not inherited and do not show up in this kind of study at all.

BRCA1 is the gene everybody in the room has heard of, and its one entry here has nothing to do with cancer. It is about the age somebody has their first period. The cancer-causing BRCA1 changes are rare, individually powerful, and found by sequencing families — a completely different method from the one that built this catalogue. So the famous cancer gene appears in a genetics catalogue under a trait nobody associates with it, because that is the only kind of variant the method could see.

The count, and even the subject of an entry, reflects which questions have been asked and with what instrument — not how variable a gene is or what it matters for. A student who takes that one idea away has learned something that applies to every dataset they will ever meet.

What the catalogue does not contain

Compare the sizes honestly. This site holds about 3,000 variants across about 2,000 genes. There are roughly 20,000 protein-coding genes, and hundreds of millions of rsIDs described. So this catalogue covers around a tenth of the genes and a tiny

fraction of one percent of the positions.

Worse than small, it is unevenly small. The studies it is built from were run disproportionately in people of European ancestry, so a position that matters in other populations is more likely to be missing — not because it does not exist, but because fewer studies went looking. That is a limitation of the source material, and it is the reason absence of a variant from any catalogue is never evidence of absence in a person.

An argument still open

Given all that, how much should anyone trust a count of variants per gene?

One position: these counts are an artefact of study design and should never be presented as a property of a gene. Another: attention is not random — a gene that keeps turning up across independent studies of different traits probably is doing something broadly important, and the count carries real signal buried in the bias.

Both are argued in the literature, and the disagreement is not settled. Students can hold the same evidence the field holds, which is the point of showing them the raw counts rather than a conclusion.

Questions

Warm up

1. In one sentence each, say what a gene is and what an rsID is.
2. Can one rsID belong to two different genes? Can one gene contain two rsIDs?
3. Roughly how many protein-coding genes does a human have, and why was the first estimate five times too high?
4. What is the double helix — the gene, the DNA, or the chromosome?

Core

1. TP53 appears once in this catalogue and PITX2 fifteen times. Give two different explanations that would both produce that pattern, and say what evidence would separate them.
2. BRCA1's only entry here is about age at first period, not cancer. Explain how the gene most famous for cancer risk ends up catalogued under something else, and what that tells you about the method used to build the catalogue.
3. A friend says "my DNA test found nothing in BRCA1, so I am fine." Explain what is wrong with that sentence, using the size comparison in this lesson.
4. Why is "a gene is a chapter and an rsID is the number given to one letter" useful, and name one specific thing it gets wrong.

5. If a catalogue is built mostly from studies in one population, what happens to the counts for genes that matter most in another? Is the resulting gap a fact about biology or about research?

To stretch

1. Design a way to test whether variant counts per gene reflect biology or attention. What would you measure, and what result would convince you?
2. Two metres of DNA fits inside a cell too small to see. Estimate the packing factor, and say what problem that packing creates for a cell that needs to read one gene.
3. rsIDs are assigned by position, and a position can be described against different reference versions of the human genome. What could go wrong when two labs compare results from different reference versions, and how would you guard against it?

Teacher notes — where each question is going

1–4 check the containment relationship and separate the physical structure (DNA, helix, chromosome) from the informational unit (gene) and the label (rsID). Question 4 catches the commonest confusion in the room: students routinely say the gene is a helix. The helix is the material the gene is written on.

5 and 6 are the centre of the lesson. Accept any two mechanisms — more studies, different study type, gene length, how variable the region is — but push hard on the second half. The evidence that separates them is the number of published studies per gene, which is countable, and students should notice that this makes the question empirical rather than a matter of opinion.

6 has a specific answer worth landing: the catalogue was built from studies of common variants, and the BRCA1 changes that cause cancer are rare ones found by sequencing families. A method that only sees common differences will catalogue a famous gene under whatever common difference it happens to carry. Students who reach this on their own have understood what an instrument is.

7 is the safety question and the one most likely to come up unprompted. The answer is that a consumer array tests a small selected panel; absence from the panel is not absence in the person. Do not let this drift into interpreting anyone's actual result.

8 rewards students who can hold a model and its limits at once. Good answers: the book is not read front to back; different cells read different chapters of the same book; most of a chapter is cut out before the instruction is used; one chapter can make several products. Students who have written code often reach for the file-and-line version in the aside — accept it, and hold them to the same standard about where it fails.

9 is about who the data was collected from. Students often assume a catalogue is a neutral mirror of nature. It is a record of what was funded and studied.

10–12 are for students who want a research problem. Question 12 has a concrete real answer — coordinates shift between genome builds while rsIDs are meant to remain stable, and mismatched builds are a genuine source of error in practice.

Take it further

Every gene and position named here has a page on this site, free and with its sources listed. Good places to continue:

- [rs1815739](#) — the ACTN3 position in the figure, one letter that changes a muscle protein.
- [Cystic fibrosis](#) — one gene, more than two thousand described variants, and a drug chosen by which one a person carries.
- [The Gene Knowledge Graph](#) — how many genes, variants and connections this catalogue holds today, and how each kind of connection is counted.
- [The papers this site is built on](#) — every study behind these pages, with every author named.

Use it freely

Free to use, copy and adapt in any classroom, at any level, including commercially. Credit is welcome and not required. Nothing on this site collects anything from a student: no account, no upload, no genetic data, ever.