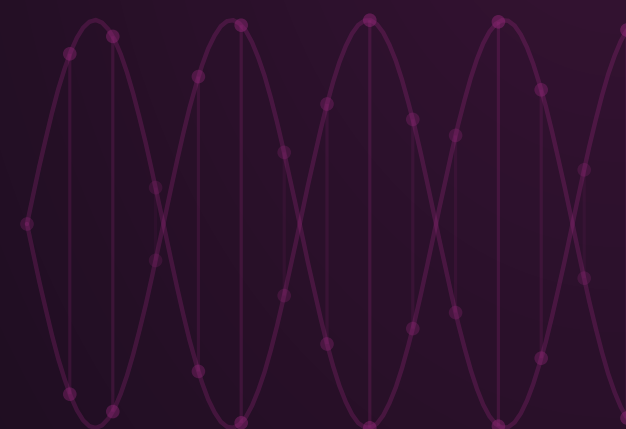


Lesson: how to read a genome file, line by line



Biology

Age 13+

Introductory

A ready-to-run introductory genetics lesson for age 13 and up, free to use in any school, college or university. What the four columns of a genome file mean, why the last one has two letters, why a position number is meaningless without knowing which edition of the map it was measured on — and the one thing such a file can never tell you. Uses our own teaching file, never a student's data.

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-reading-a-genome-file>

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: introductory. Works from about age 13 upward. No genetics needed — every word is explained here. Nothing to install, and no maths beyond dividing one number by another.
- Time: 40–50 minutes. The reading activity takes about twenty.
- You need: a browser per student or per pair, and our teaching file: `example-genotypes.txt` (15 lines, 1 KB). Print it or open it in any text editor. No genome, no sample, no test of any kind.
- It covers: what the four columns of a genome file mean, why the last one has two letters, why the position number is meaningless without knowing which map it was measured on, and the single most important thing such a file cannot tell you.

The whole thing in five lines

When someone spits in a tube and posts it off, what eventually comes back to them — under a menu called “raw data” or “download your data” — is a plain text file. Not a picture, not a report. A text file, and a fairly boring one.

- Each line is one position in the genome that the lab looked at.

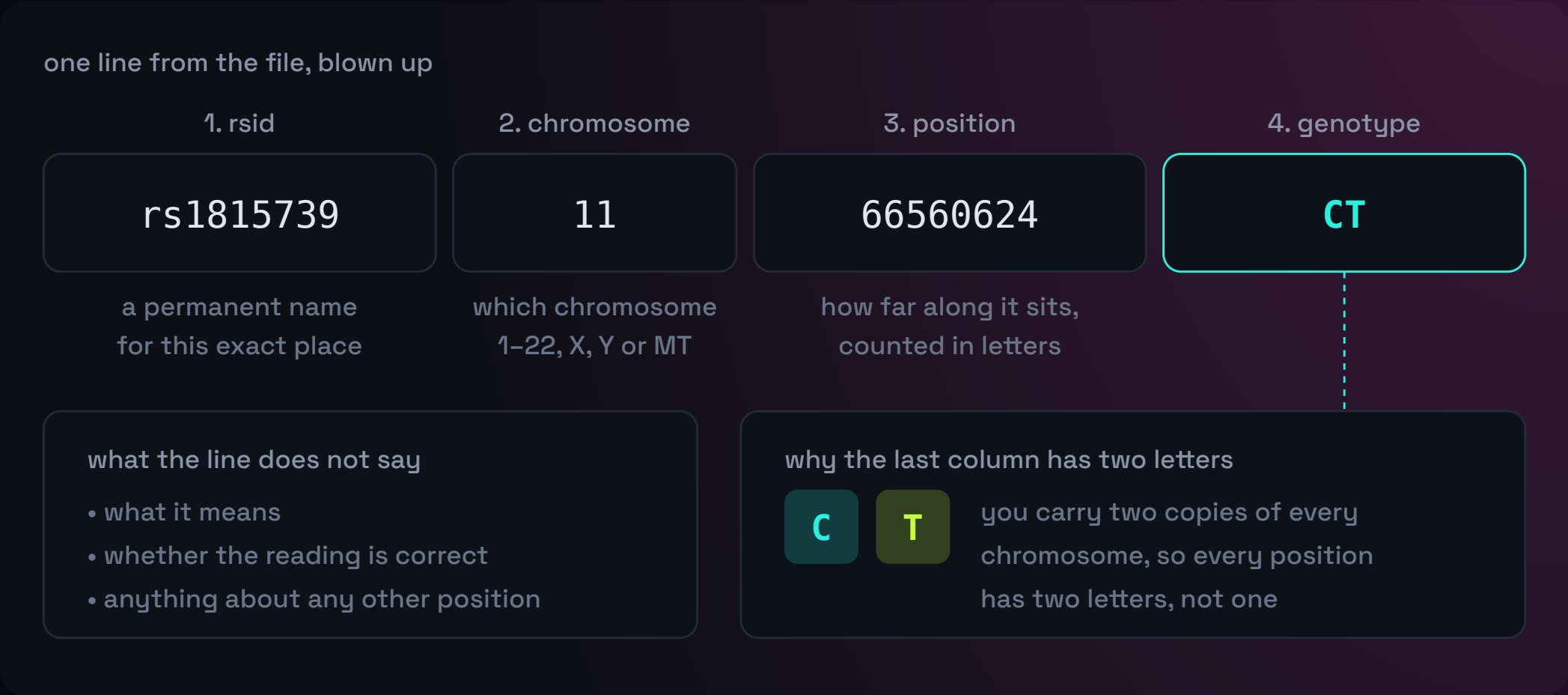
- Each line has four columns: a name for the place, which chromosome it is on, how far along that chromosome it sits, and the letters found there.
- The last column has two letters, because you carry two copies of every chromosome — one from each parent.
- The file contains no meanings. Nowhere does it say "lactose tolerant" or "high risk". It is coordinates and letters. Every meaning comes from looking the position up somewhere else — which is what this site is for.
- The file is also mostly silence. It holds a few hundred thousand positions out of roughly three billion, and the silence is not information.

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

Before you start — the one rule. Nobody brings their own genome file to this lesson, and nobody is asked whether they have ever taken a test. There is no result about any person in the room, and no list of who carries what. The file we supply is not anybody: it was generated from this site's own public catalogue, no genome was read to make it, and nothing about a person can be recovered from it. If a student asks about their own DNA, the honest answer is that a classroom is the wrong place to find out and a doctor or genetic counsellor is the right one.

One line, taken apart

Here is a single line from the file, exactly as it appears. The columns are separated by tab characters, which is why the spacing looks odd in some editors and neat in others.



Schematic. One line of the teaching file, with the four tab-separated columns pulled apart. The two letters in the last column are the two copies — one inherited from each parent. The file does not record which letter came from which parent, and nothing in the line says what any of it means.

Column	Example	What it is
rsid	rs1815739	A permanent catalogue number for one position, issued so that everybody on Earth means the same place. It names the place, never the letter anybody has there.
chromosome	11	Which of the 23 pairs the position sits on. Numbered 1 to 22, then X, Y, and MT for the small separate loop of DNA inside mitochondria.
position	66560624	How many letters along that chromosome the place sits — here, roughly 66 and a half million letters in. This number is the one with a catch, and the catch has its own section below.
genotype	CT	The letters actually found there: one on the copy inherited from one parent, one on the copy from the other. CT and TC mean the same thing, because the file does not record which came from whom.

Two identical letters — CC, GG — is called homozygous at that position. One of each —

CT — is heterozygous. Those two words are the only jargon in this lesson, and the activity below is mostly an excuse to make them concrete.

Do this: read the file

Everything here uses the teaching file and the free pages on this site. Nothing is submitted, nothing is uploaded, and no account is needed.

1. Predict first, before opening the file

Tell the class only this: the file has 15 lines, and each line’s last column holds two letters. Then ask for one number, agreed and written on the board before anybody opens anything:

Of the 15 lines, how many will show two identical letters?

Most classes answer somewhere near seven or eight. The reasoning is good and the answer is wrong, which is the useful combination. Write the prediction down; it only works if it is committed to first.

2. Count

Open the file and count. Lines beginning with # are comments and are not data — that is a convention worth pointing out, because it is the same one in almost every scientific data file a student will ever open.

What the last column shows	Name for it	Lines
Two identical letters, e.g. CC	homozygous	10
One of each, e.g. CT	heterozygous	5
Total	15	

Two the same is the common case, and here is why. At most positions the two possible letters are not equally common — one might turn up in 90 percent of copies and the other in 10. Pick two copies at random from a population like that and you will usually get two of the commoner one. Heterozygous is what you get when the two letters are closer to evenly matched, which is the rarer situation. So the shape of the column is telling you something about the whole population, not only about one person.

3. Look three lines up

In pairs, take any three lines. For each one, search this site for the rsID and write down three things:

1. Which gene, if any, the position sits in.
2. What the page says that exact genotype is associated with — not the position in general, the two letters in the file.
3. Which paper the page cites for it.

Every one of the 15 rsIDs has a page here, and every genotype in the file is one those pages actually describe, so no pair will hit a dead end. The third item matters most: a claim about a genotype is worth exactly as much as the study behind it, and students should get into the habit of asking which one it was before they are old enough to be sold anything.

4. The control: look up a line that is not there

Now give every pair the same rsID, which is not in the file: rs4244285. It has a page here. It sits in a gene called CYP2C19, and it changes how well the body activates clopidogrel, a drug given after a heart attack or a stent to stop blood clots forming. It is not a curiosity; it is the kind of thing a cardiologist wants to know.

Ask the class: what does our file say about it?

Nothing. There is no line. And now the question the whole lesson is built to reach: does that silence mean the person has nothing unusual at that position?

It does not, and it cannot. A file like this contains the positions the laboratory chose to look at — a fixed list, decided in advance, the same for every customer. A position that was never on the list produces no line, whatever the person carries there.

Absence from the file is a fact about the test, not about the person.

This is the single most misread property of these files, and it is misread in both directions: “nothing came up, so I’m fine” is wrong for the reason above, and so is “it says I have it, so I have it” — for a reason in the next section but one.

The catch in column three

The position number looks like the most solid thing on the line. It is the least.

A position is counted from one end of a chromosome, which means it depends entirely on which version of the chromosome you are counting along. The reference human genome is periodically rebuilt as gaps are closed and errors corrected, and each rebuild shifts the numbering. Two editions are in wide use at once: GRCh37, released in 2009, and GRCh38, released in 2013.

The rsID stays the same across both. The number does not.



Schematic. The same position under two editions of the reference genome. Coordinates read from Ensembl on 9 September 2026; the rsID is stable by design, the number is not.

And the shift is not a constant you could correct for. Three positions from our own file, read from Ensembl on both editions:

rsID	GRCh37 position	GRCh38 position	Difference
rs1815739	chr11 66,328,095	chr11 66,560,624	GRCh38 is 232,529 higher
rs4988235	chr2 136,608,646	chr2 135,851,076	GRCh38 is 757,570 lower
rs6025	chr1 169,519,049	chr1 169,549,811	GRCh38 is 30,762 higher

Different sizes, and not even the same direction. There is no arithmetic that converts one edition to the other; the only correct method is to look the position up again on the edition you want.

The habit to teach. The first thing to do with any file of genomic coordinates is find out which edition it was measured against — and the place to look is the header, the comment lines at the top. Our teaching file says so in its header: “Positions are GRCh38, read from Ensembl rather than typed.” If a file you are given does not say, its position numbers cannot safely be compared with anything until you find out. This is not a beginner’s precaution. Mismatched editions are a live source of error in professional work.

What the file does not contain

Get the class to do this division out loud, because the answer surprises people who have paid for one of these tests.

A human genome is about three billion letters long. A consumer genotyping file holds a few hundred thousand lines. Divide one by the other. The file covers something on the order of one hundredth of one percent of the positions in a genome — and every one of those was chosen in advance by the laboratory, because it was already known to be a place where people commonly differ.

Three consequences follow, and they are worth stating plainly to a room that will meet advertising for these products:

- Rare things are mostly invisible. The positions on the list are the common ones. Many of the changes that matter most in medicine are rare, and a test built to look at common positions will walk straight past them.
- The list reflects who was studied. The positions chosen are the ones earlier research found, and that research was run disproportionately in people of European ancestry. A position that matters more in another population is likelier to be missing — not because it does not exist, but because fewer studies went looking.
- A reading can simply be wrong. These chips make mistakes, and the mistakes concentrate exactly where they hurt: at rare variants, which are the ones a worried reader is most likely to be looking for.

That last point has been measured. A clinical laboratory in the United States reviewed 49 patient samples sent to it for confirmation of variants people had found in consumer raw data, and reported that 40 percent of those variants were false positives — the letter in the file was not the letter in the person. Several more were real but had been labelled “increased risk” when the laboratory and others classified them as benign (Tandy-Connor S and colleagues, *Genetics in Medicine* 2018, PMID:29565420).

So the honest summary of a raw data file is: an incomplete list, chosen for you, of common positions, measured with an instrument that is usually but not always right, containing no interpretation whatsoever. It is genuinely interesting. It is not a diagnosis, and no line in it should ever be acted on medically without a clinical laboratory confirming it first.

An argument still open

Given all of that, should companies hand raw data files to customers at all?

One position: they should not, or not without a clinician between. The files are routinely misread, third-party services will interpret them for a fee with no obligation to be right, and the false-positive rate above means real people have been frightened

by a letter that was never in their genome.

The other: the data is about the person, they paid for it, and withholding it is paternalism. A file people can read, question and take to a doctor is better than a polished report they must take on trust — and the way to fix misreading is to teach people to read, which is what this hour is.

Both are argued seriously, and neither has won. Students can hold the same evidence the field holds, which is why the numbers above are given raw rather than as a conclusion.

Questions

Warm up

1. Name the four columns of a genome file and say in one sentence what each holds.
2. Why does the last column have two letters instead of one?
3. A line reads GA and another reads AG. Is there any difference between them? Explain.
4. What do the lines at the top of the file beginning with # do, and why would a program reading the file need to know?
5. In the file, are there more homozygous lines or heterozygous ones? Why is that the way round it is?

Core

1. rs4244285 has a page on this site and no line in our file. Write one sentence explaining to a friend why that does not mean the person has nothing unusual there.
2. rs1815739 sits at 66,328,095 on one edition of the reference genome and 66,560,624 on another. Explain how one place can have two addresses, and say which of the two numbers is "correct".
3. Two laboratories exchange results for the same patient. One works in GRCh37 and the other in GRCh38, and neither says so. Describe a specific way this goes wrong, and one thing either laboratory could have done to prevent it.
4. A file covers roughly one hundredth of one percent of a genome, and those positions were chosen because people commonly differ there. Explain why that design makes the file good at one job and bad at another, naming both jobs.
5. Forty percent of variants sent for confirmation turned out to be false positives. Does that mean the test is 40 percent wrong overall? Explain carefully what that number does and does not measure.

To stretch

1. The file never records which parent each letter came from. Describe an experiment or a piece of extra information that would let you work it out, and say what it would let you conclude that the file alone cannot.

2. Design a header for a genomic data file that would make the errors in this lesson impossible to make. List every field you would require, and justify each one in a sentence.
3. A chip reports the letters on one particular strand of the double helix, and the two strands carry complementary letters — A pairs with T, C with G. Explain how two files could disagree about the same person's genotype while both being correct, and how you would detect it.
4. Suppose you wanted to know whether the missing positions matter. Design a study that would measure how much a consumer genotyping file misses in a population it was not designed for, and say what result would change your mind.

Teacher notes — where each question is going

1–4 check that the file has been read rather than glanced at. Question 3 has a subtlety worth drawing out: GA and AG are identical as data, and the reason is that the file never recorded which parent contributed which. Students who say “no difference” are right; students who can say why have understood something about what the file threw away.

5 is the prediction question, and the payoff is population-level. Ten of fifteen are homozygous because at most positions one letter is much commoner than the other. Some classes reach this by analogy with a bag of mostly-red counters, which is exactly the right instinct and is worth naming as such: it is the same reasoning that makes allele frequency a useful quantity.

6 is the safety question and the one most likely to arrive unprompted, often phrased as “my test said I was clear”. Land it firmly: the list of tested positions is fixed in advance and identical for every customer, so absence from the file is a property of the test. Do not let this drift into interpreting any actual result belonging to anyone in the room.

7 has an answer students often resist: neither number is correct in isolation, because a coordinate is meaningless without the edition it belongs to. The rsID is the stable identifier and the coordinate is a derived value — which is precisely why rsIDs exist. Good answers notice that this is the same reason a street address needs a city.

8 is the professional version of question 7 and rewards concreteness. Accept any specific failure — a variant looked up at the wrong position and reported as absent, two datasets merged and silently misaligned, a coordinate that lands inside a different gene entirely. The prevention is a mandatory header field, which sets up question 12.

9 wants both halves. Good at: measuring common differences cheaply across very many people, which is what made large population studies possible at all. Bad at: telling one individual whether they carry a rare, serious change — the

exact use most customers have in mind.

10 is a statistics question hiding in a genetics lesson, and it is the hardest one here. The 40 percent is a rate among variants that somebody thought worth sending to a clinical laboratory — a heavily selected group, skewed towards rare and alarming results, which is where these chips are weakest. It does not describe the file as a whole. Students who reach “the denominator is not what you think” have learned something that will serve them far outside this subject.

11–14 are for students who want a research problem. Question 11 points at phasing, and the everyday answer — test a parent — is a good one. Question 13 is the strand problem, which is real, catches professionals, and is detectable because an A/G position reported as C/T is complementary rather than merely different; the genuinely ambiguous cases are A/T and C/G, where flipping the strand gives you back a valid-looking answer with no way to tell. Students who find that on their own have found the thing that makes the problem hard.

A note on the file itself. It is not a plausible person and is not meant to be: the genotypes were chosen to give the exercise one of each case, not sampled from any population, so several rare combinations appear together in a way they never would in life. That is worth saying to the class if anyone notices — and a student who notices has understood the previous section better than the exercise required.

Take it further

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

- Genes and rsIDs, and what a catalogue measures — the companion lesson: where these names come from, and why counting them measures research attention rather than biology.
- rs1815739 — the line taken apart in the figure above, one letter that changes a muscle protein.
- rs4244285 — the control: the position our file is silent about, and a drug whose dose depends on it.
- The Gene Knowledge Graph — how many positions, genes and connections this catalogue holds today, counted live.
- 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.