Case study: The distribution of human height

Preliminaries

If you are not already familiar with the structure of these exercises, read the Introduction first.

> **Note**
>
> Reminder: Save your work regularly.

> **Important**
>
> If you are using a Mac, we recommend that you use either Chrome or Firefox to complete these exercises. Some of the default settings in Safari prevent these exercises from running.

Contact information

If you have questions about these exercises, please contact Dr. Kevin Middleton (middletonk@missouri.edu) or drop by Tucker 224.

Learning objectives

The learning objectives for this exercise are:

- Contrast polygenic traits with Mendelian traits
- Demonstrate how quantitative traits with continuous-valued phenotypic measures result from the combined effects of many different genes
- Describe the process by which many genes can each contribute a small amount to a measurable phenotype

Quantitative traits result from combinations of many alleles

So far we have built up from simple Mendelian traits to distributions of many alleles taking on the shape of a normal distribution. How can we make the leap from combinations of alleles to quantitative traits?

The solution is to assign a small positive or negative value to each allele, and the size of that value depends on many factors. In essence, we can imagine that any quantitative trait has a baseline value which is modified up or down by the

presence of an allele. By counting the numbers of “positive” and “negative” alleles, we can arrive at a phenotypic measurement.

To do this, we have to make some assumptions:

- All genes have roughly equal effects (no gene has more impact on the phenotype than any others).
- All genes act additively, so that we can count alleles to arrive at a phenotype. Additivity can mean adding negative numbers though.
- Genes do not interact with one another (*epistasis*) or with the environment (*genotype by environment* interactions).
- Our simulation accounts for all of the phenotypic variation in a trait.

In real world biological systems, none of these assumptions is completely met to one degree or another. Nonetheless, we can use this framework to begin to understand quantitative traits.

Alleles to quantitative traits

Let’s return to the allelic combination plot for 5 genes (Figure 1). There are more than 1,000 possible combinations of alleles, but only 10 possible resulting genotypes, from 0A/10T to 10A/0T. The most likely combination is 5 A and 5 T.

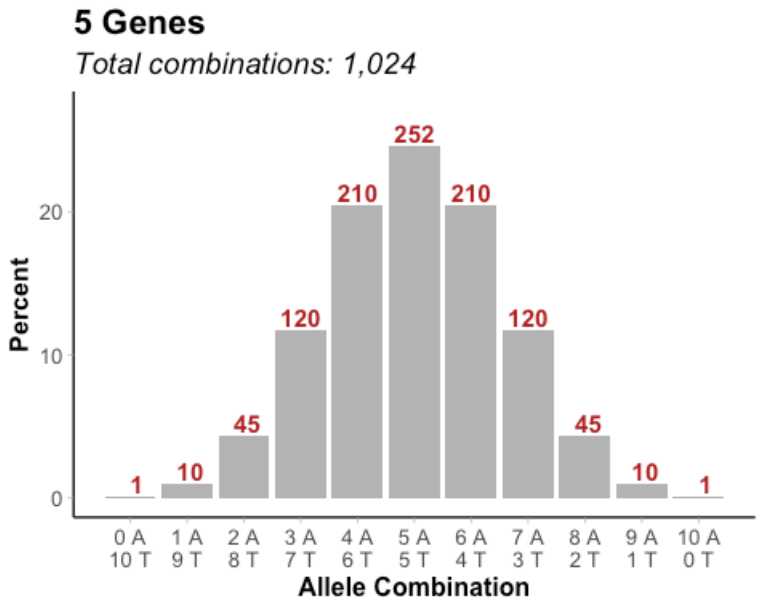


Figure 1: Allelic combination plot for 5 genes.

For this simulation “experiment,” we will use a common plant mode organism: *Arabidopsis thaliana* (a relative of mustard; Figure 2). *Arabidopsis* is commonly used to study the genetics of quantitative traits in plants, because it grows quickly

and easily in a greenhouse, where the environment can be easily controlled for experiments.

Figure 2: Arabidopsis thaliana, a common model organism for plant quantitative genetics.

A common measure of the amount of growth in plants is above ground biomass. Plants, such as *Arabidopsis* are allowed to grow in controlled conditions. After a certain period of time, the plants are harvested, dried, and weighed.

Imagine that under certain conditions, the mean above ground biomass of *Arabidopsis* is 5000 mg (i.e., 5 g) and that 5 genes control the range of biomass. Each T that a plant receives results in 50 mg lower biomass, and each A that it receives results in 50 mg higher biomass.

A plant with 10 T and no A would weigh 5000 - (10 * 50) = 4500 mg. Each of the 10 T's subtracts 50 mg from 5000. Similarly, a plant with 10 A's would weight 5500 mg.

What would a plant with 5T and 5A weigh? Briefly explain your reasoning.

5000 mg. The 5T and 5A would all cancel each other out, leaving the "default" above ground biomass.

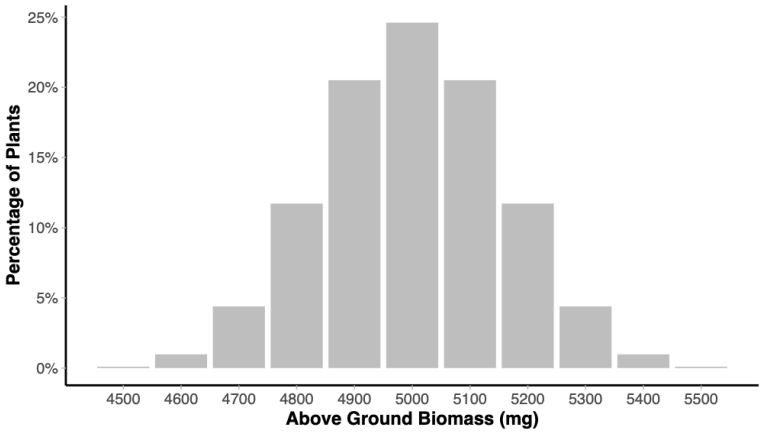
If we randomly sampled *Arabidopsis*, we would expect to find plants with genotypes matching the expected distribution in Figure 1.

What do you predict the range of above ground biomass in Arabidopsis would be?

4500 to 5500 would be the range from all T to all A.

If we were to weigh a large number of *Arabidopsis* plants, we would find a distribution that looks like the one produced in the chunk below. Run the code to make the plot.

```
biomass_plot()
```



There would be plants with exactly 4500 mg, 4600 mg, 4700 mg, etc. of above ground biomass. If each allele only changes biomass by 50 mg, why do we not find plants that weigh 4550 mg or 4650 mg?

We set up the simulation such that A and T either add or subtract 50 mg. One less A means one more T. So the distribution has to go in steps of 100 mg.

Our simulated set of plant biomasses reveals a major limitation of our approach:

- Why would plants weigh exactly 4500 mg or 4600 mg or 4700 mg with no plants at intermediate weights?

The answers lie in the assumptions that we made at the start. In a simple simulation experiment like this, we have to make simplifying assumptions that are often not realistic in real biological systems. We could, however, make our model complex to better approximate the natural world.[1]

Let's explore a set of data to begin to see how biologists study quantitative traits in the real world.

Case study: the distribution of human height

The National Health and Nutrition Examination Survey ("NHANES") began in the early 1960s and continues to the present time. The goal of this study is to assess the health and nutrition status of a broad cross-section of the United States population. As part of this study, routine measurements of body size such as height (in cm) are recorded for each participant.

---

[1] Such simulation models with increasing levels of complexity are quite common in the field of genetics.

The 2017-2020 NHANES survey has data for 13,137 individuals. Figure 3 shows the observed heights for all the individuals in the study. Both groups show a similar pattern of roughly linear increase in height from childhood until age 15-18. What follows is a slight decline in height as the spaces between the intervertebral discs decrease slightly.

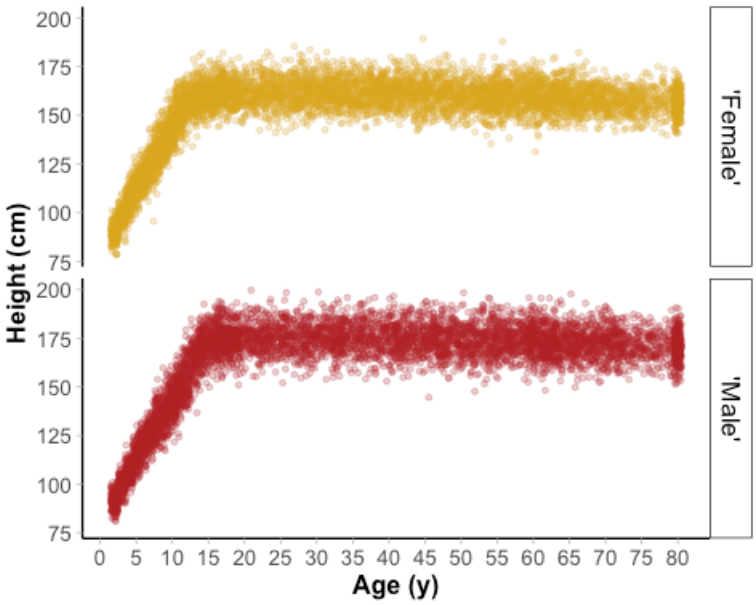


Figure 3: Heights for NHANES participants who either self-identify as Female or were assigned female at birth ('Female') or who self-identify as Male or were assigned male at birth ('Male'). These broad categorizations mask extensive variability in the human population. Recent estimates suggest that at least 1 in 5,000 humans are intersex, with some estimates as high as 1 in 1,000. Data from NHANES 2017-2020.

If we ignore the growth phase by selecting individuals over age 20, we can get a reasonable sample of adult heights. Figure 4 shows the distributions of heights for both groups. We can see that the range is between about 140 and 200 cm, but the majority of individuals fall near the middle of that range.

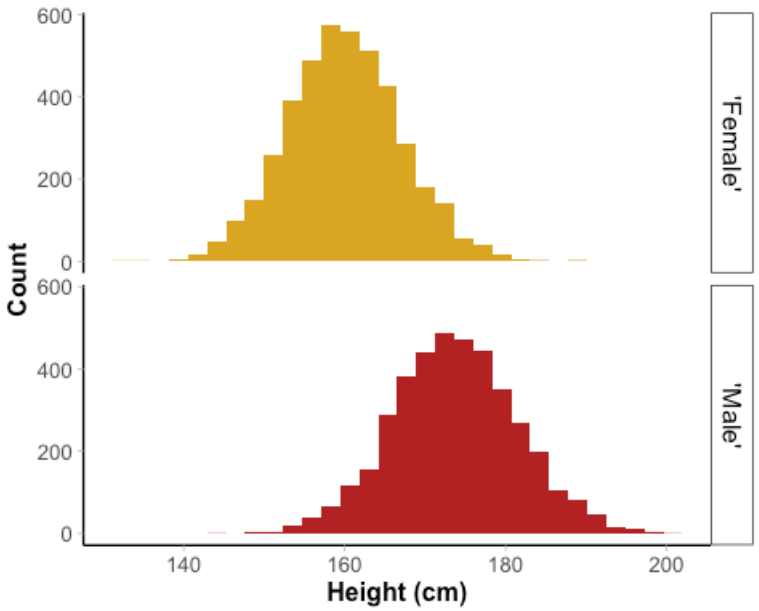


Figure 4: Distributions of height for all individuals in the NHANES study over age 20. Data from NHANES 2017-2020.

Generating a normal distribution from combinations of alleles

For the remainder of this exercise, we will use the data in yellow. Using what we have learned about distributions of alleles, we want to explore how many genes might be responsible for the variation in height that we observe.

Figure 5 shows the distribution of observed heights (4,267 individuals over the age of 20) in the upper panel. The lower panel shows the results of a simulation where the observed variation in height is distributed among 5 genes (10 alleles). Each allele has to account for about 2.8 cm of height to account for >40 cm of range in height.

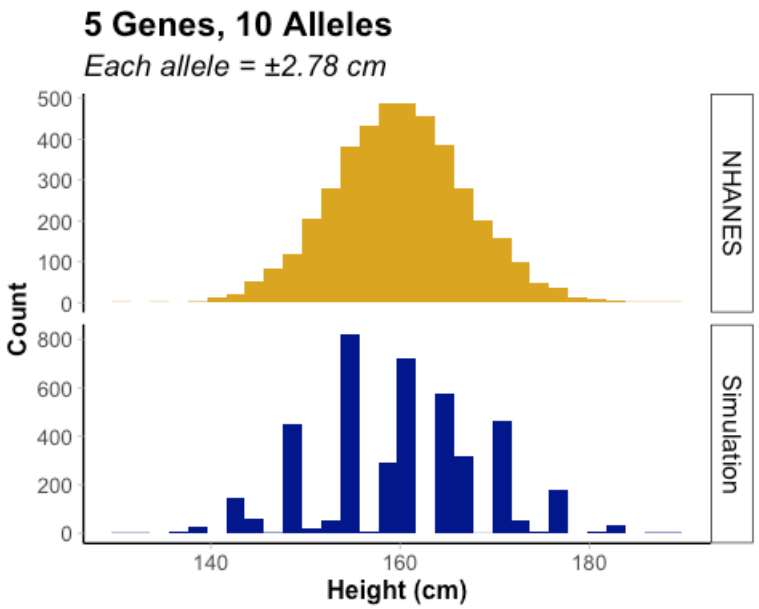
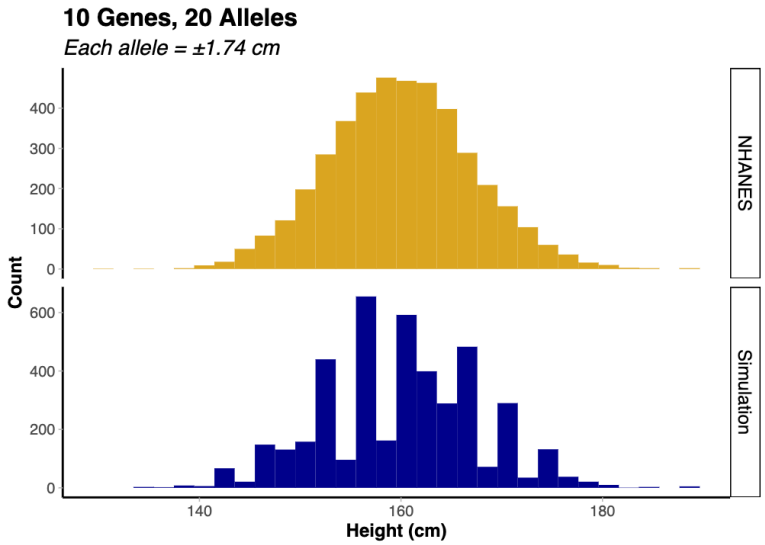
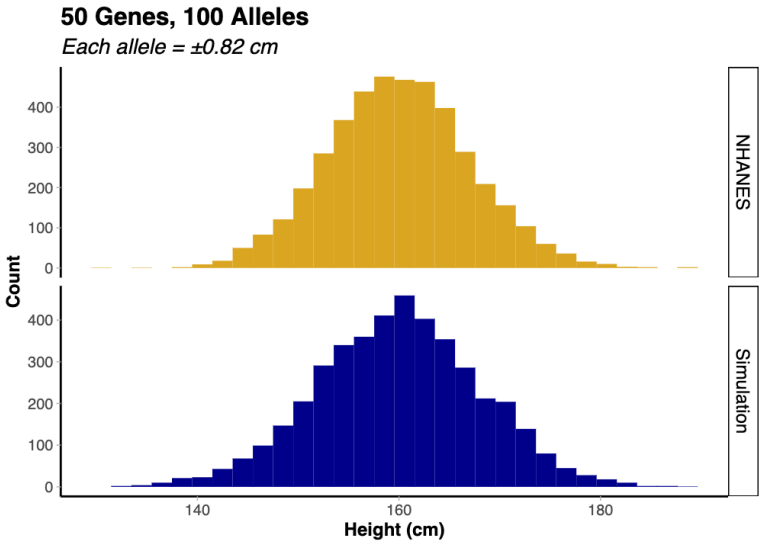


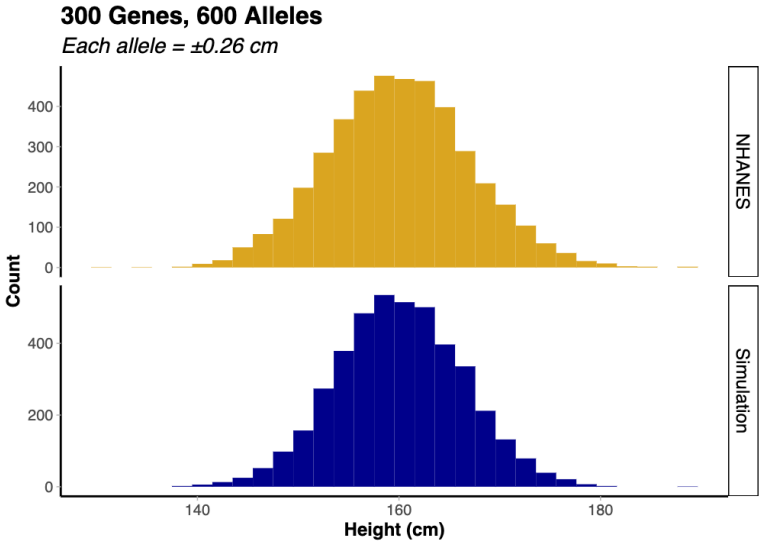
Figure 5: Observed heights (upper panel) and simulated heights (lower panel) for five genes with the observed variation divided among each of the genes. The breaks in between the histogram bars result from the relatively low number of contributing genes, each of which must account for a large proportion of the variation in height.

Using the code block below, try increasing numbers of genes (e.g., 10, 20, 50, 100, 300, etc.).

```
simulate_heights(n_genes = 10)
```







As the number of genes increases, how do the distributions of actual heights and simulated heights compare to one another? How does the amount of phenotypic variation attributable to each allele change as the number of genes contributing to height increases?

As the number of genes increases, the distributions appear more and more similar. With 300 genes, they are almost identical. Simultaneously, the amount of phenotypic variation decreases from ~2.8 cm to ~0.26 cm.

Summarizing distributions

Many biological traits, including those related to size (length, height, mass, etc.), result from the actions of large numbers of genes, each adding or subtracting a

small amount of a phenotype. Because of the actions of a large number of genes, these traits often follow a normal distribution.

Biologists are very often interested in summarizing a set of observations (*sample*). Two numbers are all that are needed to fully describe a normal distribution: the *mean* and the *standard deviation*. You are probably already familiar with the mean (often called the average).

The mean ($\bar{y}$; the bar over y denotes a mean) is the sum of the observed values (y) divided by the number of observations (n):

$$\bar{y} = \frac{\sum(y)}{n}$$

The standard deviation (s) is a little more complicated:

$$s = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}$$

The standard deviation involves the squared deviations of each observed value (y) from the mean ($\bar{y}$) divided by the number of observations minus 1 ($n - 1$), with the square root taken of those values.

Think of the standard deviation as a measure of how far, on average, each point falls from the mean.

There are built-in functions to do these calculations for us, so we don't have to keep track of all those deviations. Run the code chunks below to find the mean and standard deviation of the heights data.

```
# Mean
mean(NHANES$Height)
```

```
[1] 160.0093
```

```
# Standard deviation
sd(NHANES$Height)
```

```
[1] 7.070795
```

The mean is about 160 cm, and the standard deviation is about 7.1 cm. We can use these two numbers to define a normal curve for these data, because the shape of the normal distribution only depends on these two numbers (Figure 6).

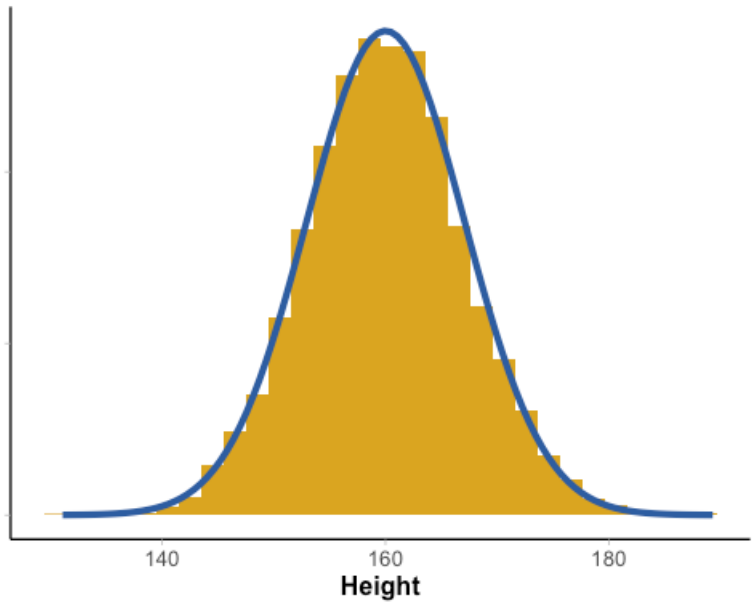


Figure 6: Histogram of observed heights (cm) for 4,267 individuals (yellow bars). The blue line represents a normal distribution with the same mean and standard deviation as the observed data. The two overlap almost perfectly. y-axis labels are omitted, because the histogram and distribution have different scales.

Working with distributions

One of the features of a normal distribution is that we can use the mean and standard deviation to tell us about the range in which we expect to find most of the observations.

As one example, in a large sample that is normally distributed, we expect that 95% of the observations will fall within about 2 standard deviations (more accurately 1.96 standard deviations) of the mean.

The code chunks below do this calculation for you and save the lower and upper bounds of this interval into two new variables.

```
# Lower bound
(lower <- mean(NHANES$Height) - 1.96 * sd(NHANES$Height))
```

```
[1] 146.1505
```

```
# Upper bound
(upper <- mean(NHANES$Height) + 1.96 * sd(NHANES$Height))
```

```
[1] 173.868
```

We can use these numbers to determine whether our observed sample is normally distributed in reality.

In the code chunks below, first first we record the number of observations and save it to a variable (`n_observations`). We then count the observations that are below the lower bound. This count is then divided by the number of observations and multiplied by 100 to determine the percentage of observations below the predicted lower bound. This process is repeated for the upper bound

```
n_observations <- length(NHANES$Height)

# Below 146.1505 (lower)
sum(NHANES$Height <= lower) / n_observations * 100

# Above 173.868 (upper)
sum(NHANES$Height >= upper) / n_observations * 100
```

```
[1] 2.531052
```

Based on the predictions above and the percentages you calculated, does it appear that heights are normally distributed in this sample? Why or why not?

They are very close to normal. About 2.53% of the individuals are lower than the lower bound. About 2.60% are above. If anything, there are a few more tall individuals than we would expect (in this sample).

Finally, we can look at the individuals with the most extreme heights. These are 131.1 cm and 189.3 cm.

```
# Lowest
head(NHANES)
```

```
     Age  Height
3406  60   131.1
2993  54   135.3
7653  57   138.3
1065  61   139.0
4409  54   139.7
9543  79   139.7
```

```
# Highest
tail(NHANES)
```

```
     Age  Height
3176  23   182.4
7330  31   182.4
1350  29   183.7
```

```
8624  27   185.3
9728  56   187.8
1532  45   189.3
```

Epilogue: Mapping the genes for height in humans

To this point, we have only considered how genes contribute to a quantitative trait and how many genes might contribute to a trait. What we haven't considered yet is how scientists estimate where in the genome the associated genes are located ("mapping"). Many different methods are used for mapping. Since the beginning of the genomic era, where full genomes can be sequenced and compared, one of the most common methods for mapping is via Genome-Wide Association Studies (GWAS).[2]

GWAS compares large numbers (hundreds of thousands or millions) of single-nucleotide polymorphisms (SNPs) to their associated phenotypes through large-scale statistical testing. These tests reveal locations on the genome that correspond to measurable phenotypic variation.

Genes contributing to human height

Because it is easily measured during routine visits to a physician, human height is one of the best studied quantitative traits. It was the focus of many of the largest early genomic studies. One such study was by Lango Allen and colleagues (2010). These authors identified approximately 700 genes that together were able to explain about 16% of the variation in height (Figure 7). Remember that our simulation above was designed to account for 100% of the phenotypic variation.

[2] Uffelmann and colleagues (2021) present an introduction.

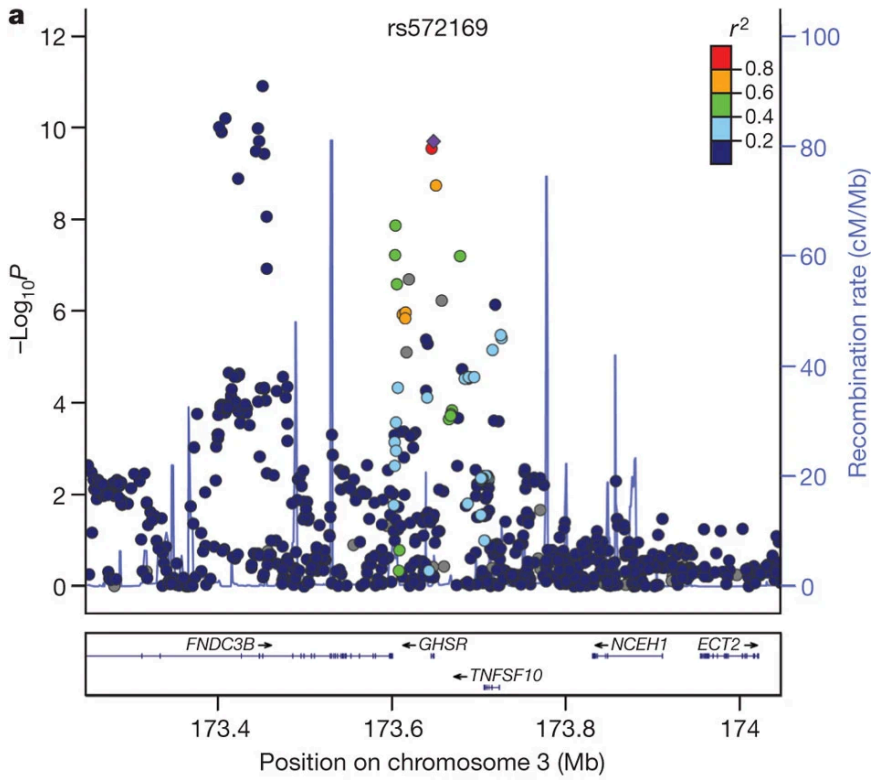


Figure 7: Partial results of a GWAS investigating the genes involved in human height. A small section of human chromosome 3 shows a region of high association between height and single nucleotide polymorphisms. The points with high values on the -$Log_{10}P$ axis indicate a significant association between that SNP and height. This region of the genome contains the gene GHSR, which codes for the Growth Hormone Secretagogue Receptor. GHSR is thought to be involved in the body's energy regulation system. Image from Lango Allen et al. (2010).

More recently, Yengo et al. (2018) described the largest GWAS to date for human height. This study included genetic data for ~700,000 individuals of primarily European descent. These authors identified 3,290 SNPs that collectively explain about 25% of the phenotypic variation in human height.

Even the largest GWAS study leaves 75% of the variation currently unexplained.

Feedback

We would appreciate your anonymous feedback on this exercise. If you choose to, please fill out this optional 4-question survey to help us improve.

References

Lango Allen, H., and many others. 2010. Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height. *Nature* 467:832–838.

Uffelmann, E., Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, and T. Lappalainen. 2021. Genome-Wide Association Studies. *Nature Reviews Methods Primers* 1:1–21. Nature Publishing Group.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, M. N. Weedon, T. M. Frayling, J. Hirschhorn, J. Yang, P. M. Visscher, and GIANT Consortium. 2018. Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ~700000 Individuals of European Ancestry. *Hum. Mol. Genet.* 27:3641–3649. Oxford University Press (OUP).