Introduction
- Fission yeast
- GWAS for yeast
- The traits

What does this data mean?

How does this influence evolution?
- Another trait: glucose & fructose use.
- Linked alleles
- Some more things to think about and discuss
- Summary

BIO00056I GWAS Workshop

Introduction

This is GWAS workshop for Genes & genomes in populations & evolution - BIO00056I.

In the lecture last week, we learnt about quantitative traits and how to analyse them. To recap, quantitative traits are often described as "complex" traits, because they are usually controlled by many genes, which each account for a small proportion of the trait variation. For this reason, they are known as additive genes.

As we saw in the lecture, one good way to identify loci controlling quantitative traits is to perform a Genome-Wide Association Study, or GWAS. GWAS experiments use diversity panels of individuals to identify genetic markers that are significantly associated with a trait of interest.

Using a diversity panel instead of a biparental mapping population ensures that linkage disequilibrium is low (bits of co-inherited DNA are small), and therefore resolution is high (QTL sizes are small), which helps to narrow down the regions containing the genes controlling the trait. They also capture much more genetic variation.

Fission yeast

In this workshop we'll use data from the fission yeast *Schizosacharomyces pombe*, which is a model species for cell biology. You can read more here about *S. pombe* here (https://en.wikipedia.org/wiki/Schizosaccharomyces_pombe), and there are some simple facts about the genome here:

Genome facts
12 Mb genome
~5,000 protein coding genes
~1500 non-coding RNAs
Haploid

When fission yeast is not being used in labs as a model species for cell biology, it grows on fruit. It is most often found in high-sugar fruits or in fermented beverages (wine, beer, and quite often in cachaça in Brazil).

The data we will use is from two papers from Nature Genetics (https://www.ncbi.nlm.nih.gov/pubmed/25665008), and Nature Communications (http://europepmc.org/abstract/MED/28117401). These articles described the genetic diversity of 160 strains of this species by sequencing their genomes. We also measured over 200 traits.

GWAS for yeast

GWAS is usually used for human or crop traits, but of course we can apply this method to any recombing sexually-recombining eukaryote species where we have trait data and polymorphism data (eg: SNPs).

In this workshop we'll have a look at the GWAS results for four traits, and think about what these results could mean for the evolution of this species.

The traits

The traits we will look at are:

Growth in 400mM NaCl This trait is the growth rate of the strains in 400mM salt. This could be important if strains grew in fruit near the sea, for example.

Growth at an elevated temperature This trait is the growth rate of the strains at 40 degrees centigrade. This is quite warm for this yeast, so it's a stressful condition.

The maximal cell density in rich media This trait measures how well each strain grows, if it's given plenty of sugar and other resources.

The glucose/fructose utlisation level in wine This trait measures how much of the glucose and fructose sugars each strain used in grape must (which is used to make wine).

A high value in any of these traits could be adaptive one condition or another. But optimising one trait may limit another, particularly if the alleles that influence two traits are closely *linked*.

If you have touble with any of the R code, you can download it all here (http://www-users.york.ac.uk/~dj757/BIO00056I/misc/gwas-code.R)

Exploring the data

First, clear all the previous data you may have in R.

```
rm(list = ls())
```

Then set your working directory, either using the session menu, or a command
.

Now, load all the GWAS data in to R

You can load the data we will use directly from the URL like this...

```
load(url("http://www-users.york.ac.uk/~dj757/BIO00056I/data/gwas-data.Rda"))
```

After loading this data you'll then have four data frames of data called **gluc**, **heat**, **rich** and **salt**. Each of these data frames contains the GWAS results for one trait. Let's start by looking at one table, salt. Try these commands:

```
head(salt)
nrow(salt)
summary(salt)
View(salt)
```

This table contains the output from GWAS program called LDAK (http://dougspeed.com/ldak/). It describes the association of each SNP with the trait of salt tolerance.

Each row of the table contains information about a SNP. For our purposes, the important columns are **Wald_P** (the p-value), **Chr** (which chromosome the SNP is on), **BP** (the position in the chromosome) and the **Effect_size** (the effect size of the variant). The **effect size** of the variant estimates the proportion of trait variation that is explained by this SNP.

Make a Manhattan plot

A Manhattan plot is a standard way of viewing GWAS results. They are named Manhattan plots because they resemble the skyline of a highrise city, such as Manhattan Island (New York).

Manhattan plots show the position of the SNP on the *x* axis, and the p-value of the association on the *y* axis. P-alues are plotted as -log10(P), so that very small p-values end up at the top of the plot, and stand out.

The code below will make a Manhattan plot for the salt trait

```
#set plotting parameters
layout(matrix(c(1:3), 1, 3, byrow = TRUE),widths = c(5579133,4539804,245288
3))
par(bty="l",cex=0.8,mar=c(2,2,2,0)+2)
#loop through chromosomes 1,2,3, plotting each one in turn
for (j in 1:3){
      #make a subset of SNPs from this chromosome
      temp = subset(salt, Chr == j)
      #plot these
      plot(temp$BP, -log10(temp$Wald_P),
           main = paste("Chr",j),ylim=c(1,10),
           xlab="position",ylab="-log10(P)"
      )
      #draw a threshold line
      #SNPs above this line are statistically associated with the trait
      abline(h=-log10(1e-5),col=2,lty=2)
}
```

What does this data mean?

This Manhattan plot takes the -log10 transformed P-values for every marker-trait association test (Y-axis), and plots them against their position in the genome (X-axis). *S. pombe* has three chromosomes, so we have mapped them onto three separate plots.

The p-value transformation means that the best P-values (which are very small) are now the highest points on the plot. The red lines are a significance threshold. Markers below the line are considered non-significant, and those above the line are significantly associated with the trait.

The strength of the association between the genotype and the phenotype is related to the **effect size** of the variant (how strongly it affects the trait). Some alleles will affect the trait very strongly, others not so much. Now let's look at how large the effect sizes are, by making a histogram of all the effect sizes.

```
hist(salt$Effect_size,br=30)
```

Consider carefully what this histogram tells you about SNP effect sizes.

How does this influence evolution?

The **effect size** of a variant is *the proportion of the phenotypic variance that the variant explains*. We can see that most of the effect sizes are very small. Most are smaller than 0.05, meaning that they produce an effect that is smaller than 5% of the trait.

At this point let's think about how many small-effect alleles might affect evolutionary change.

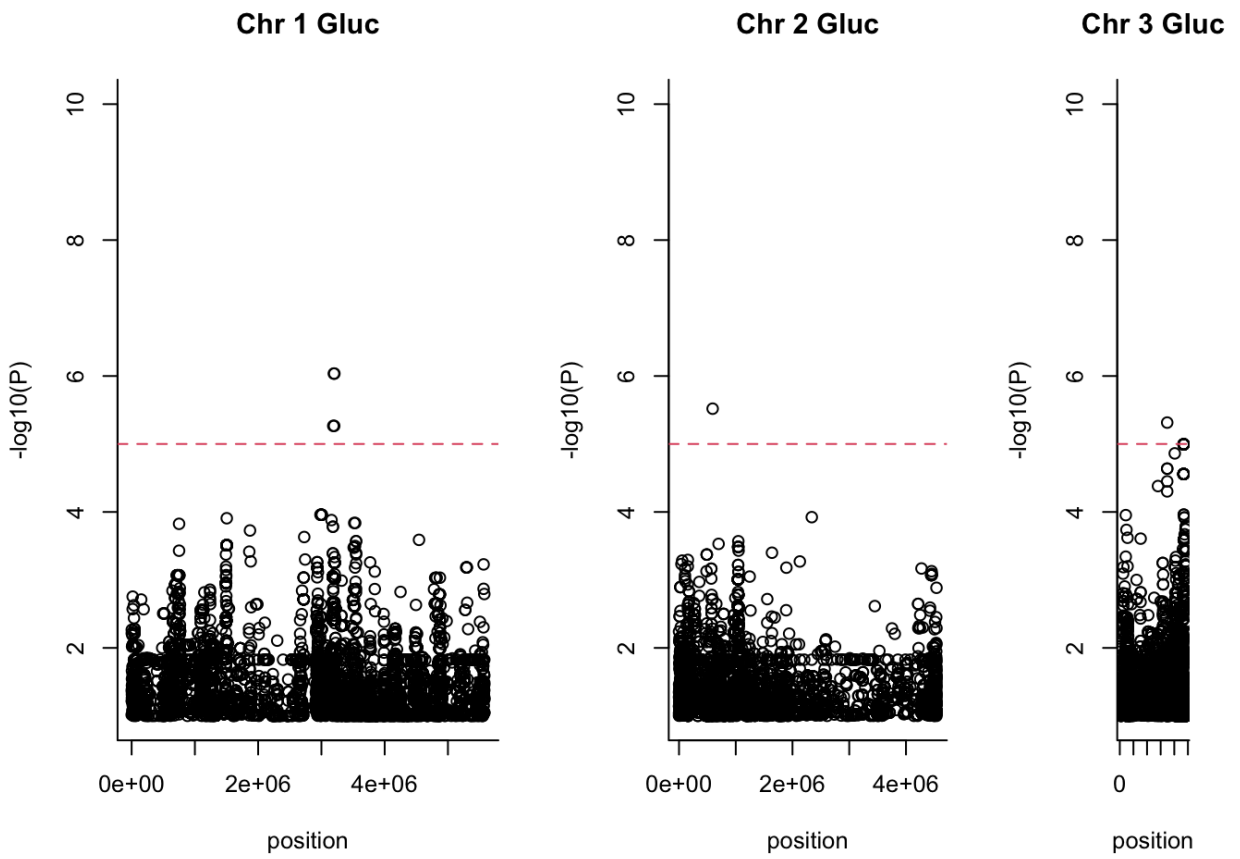
A selective sweep is when a beneficial mutation becomes fixed in a population * Would we expect to observe one sweep of evolution to fix any of these small effect variants? * Would we expect this to occur if they fall close to very deleterious alleles?

Another trait: glucose & fructose use.

In a vineyard environment, where this yeast grows, it will be an advantage to be able to metabolise the glucose & fructose in the rotting fruit. One of the other traits in our dataset is a measure of the glucose & fructose use. Lets have a look at that. This data is in the **gluc** data frame.

```
layout(matrix(c(1:3), 1, 3, byrow = TRUE),widths = c(5579133,4539804,245288
3))
par(bty="l",cex=0.8,mar=c(2,2,2,0)+2)

for (j in 1:3){
      temp = subset(gluc, Chr == j)
      plot(temp$BP, -log10(temp$Wald_P),
           main = paste("Chr", j, "Gluc"),ylim=c(1,10),
           xlab="position",ylab="-log10(P)"
      )
      abline(h=-log10(1e-5),col=2,lty=2)
}
```



Again, there are some very low P-values and many low ones, suggesting many small-effect alleles. Plot a histogram of the effect sizes, as before, like so:

```
    hist(gluc$Effect_size)
```

Now let's compare the two Manhattan plots by plotting them on top of each other. We will plot the salt results in black, and the glucose metabolism results in red, and write the plot output to a pdf file. This pdf will appear in your working directory.

```
#open the pdf writer
pdf("GWAS-plot3.pdf",width=60)
#set up the plot layout
layout(matrix(c(1:3), 1, 3, byrow = TRUE),widths = c(5579133,4539804,245288
3))
par(bty="l",cex=0.8,mar=c(2,2,2,0)+2)
#loop through chromosomes
for (j in 1:3){
      #plot the salt GWAS points (black dots)
      temp = subset(salt, Chr == j)
      plot(temp$BP, -log10(temp$Wald_P),main = "",ylim=c(1,10),xlab="positio
n",ylab="-log10(P)")

      #add the glucose/fructose GWAS points (red crosses)
      temp2 = subset(gluc, Chr == j)
      points(temp2$BP, -log10(temp2$Wald_P),main = "gluc",ylim=c(1,10),xlab
="position",ylab="-log10(P)",col=2,pch=3)

      #add the significance threshold
      abline(h=-log10(1e-5),col=2,lty=2)
}
#close the pdf writer
dev.off()
```

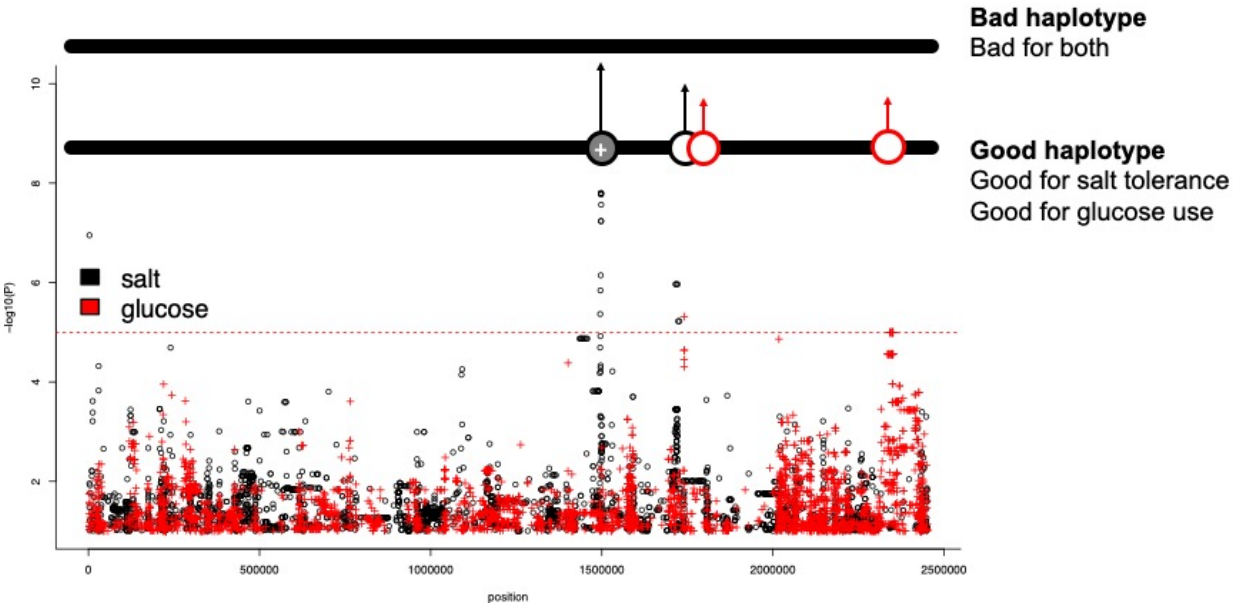
Now find your pdf, open it, and look at the plot.

Linked alleles

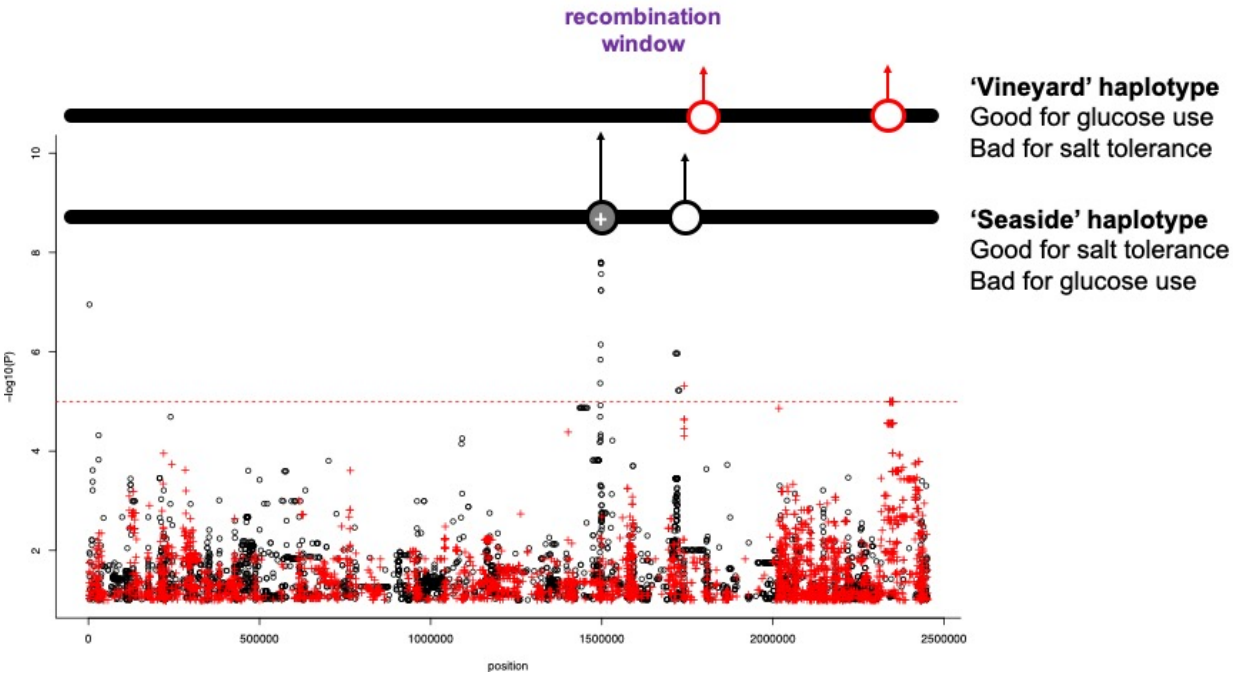
Look at that significant peak on chromosome 3, at position ~1,500,000. Remember, this will have a **large effect size**. See how there is also an important variant for glucose & fructose use on chromosome 3, near the same position.

Let's consider what might happen if alleles are **linked** and the environment is complex (which all environments are). In this case, we consider an environment which is salty, but has plenty of fructose. A coastal vineyard, for example.

- What might happen if the all the helpful alleles are on the same haplotype? Like so:



- What might happen if the two helpful alleles are on different (opposing) haplotypes?



Some more things to think about and discuss

Use the script above to plot some other pairs of traits. Can you find any other examples where peaks for different traits are found in close proximity to each other? What implications could this have for *S. pombe* or its use?

Are all of the significant markers in peaks? Apart from being close to a QTL for the trait of interest, What else could lead to strong P-values?

Which trait has the highest association peaks? What are the factors that affect this?

Summary

Natural populations contain many thousands of segregating alleles. Even this small sample of 160 strains, in this small yeast genome contained 173,000 SNPs, 14,000 small indels and 1,000 long terminal repeat (LTR) insertions from retrotransposons, see here (https://www.nature.com/articles/ng.3215).

Any of these alleles can have small or large effects on phenotypes. We have seen today that they are mostly very small (this is consistent with a great deal of other research).

Whether these alleles become more frequent, or are lost from a population depends on many things, including:

- the environment (seaside or vineyard in our example)
- the effect size of the allele
- what other alleles are on the haplotype
- the recombination rate
