Answers_Association_exercise.Rmd

Answers

This document contains answers to the exercise that accompanies **Microbiome association analysis** lecture. R commands are included where relevant.

Q1: 393 individuals and 7381 MGS+CAG's.

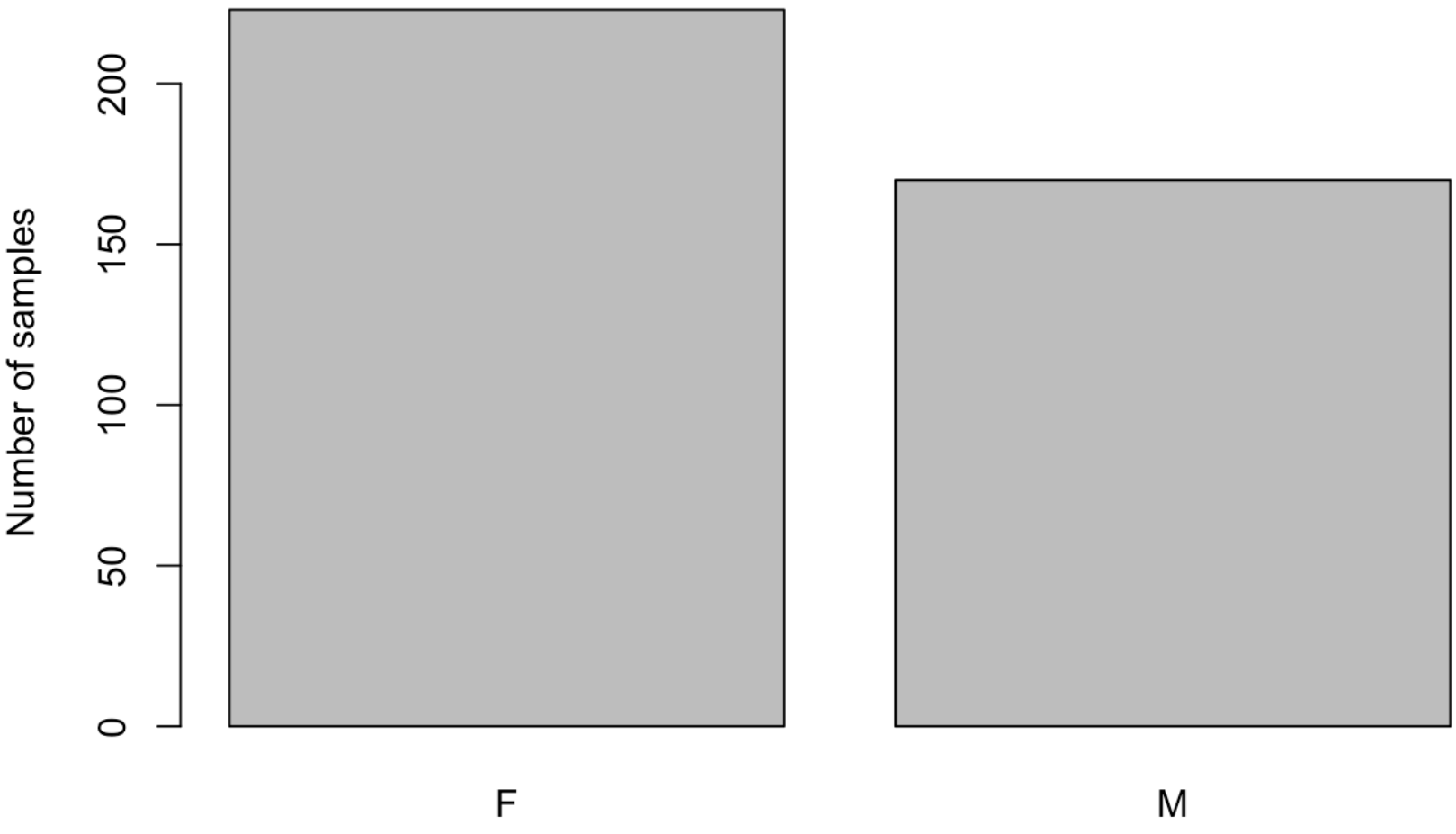
```
str(speciec_matrix)
```

Q2: Grep searches for a specified pattern and returns an index of elements that match the pattern.

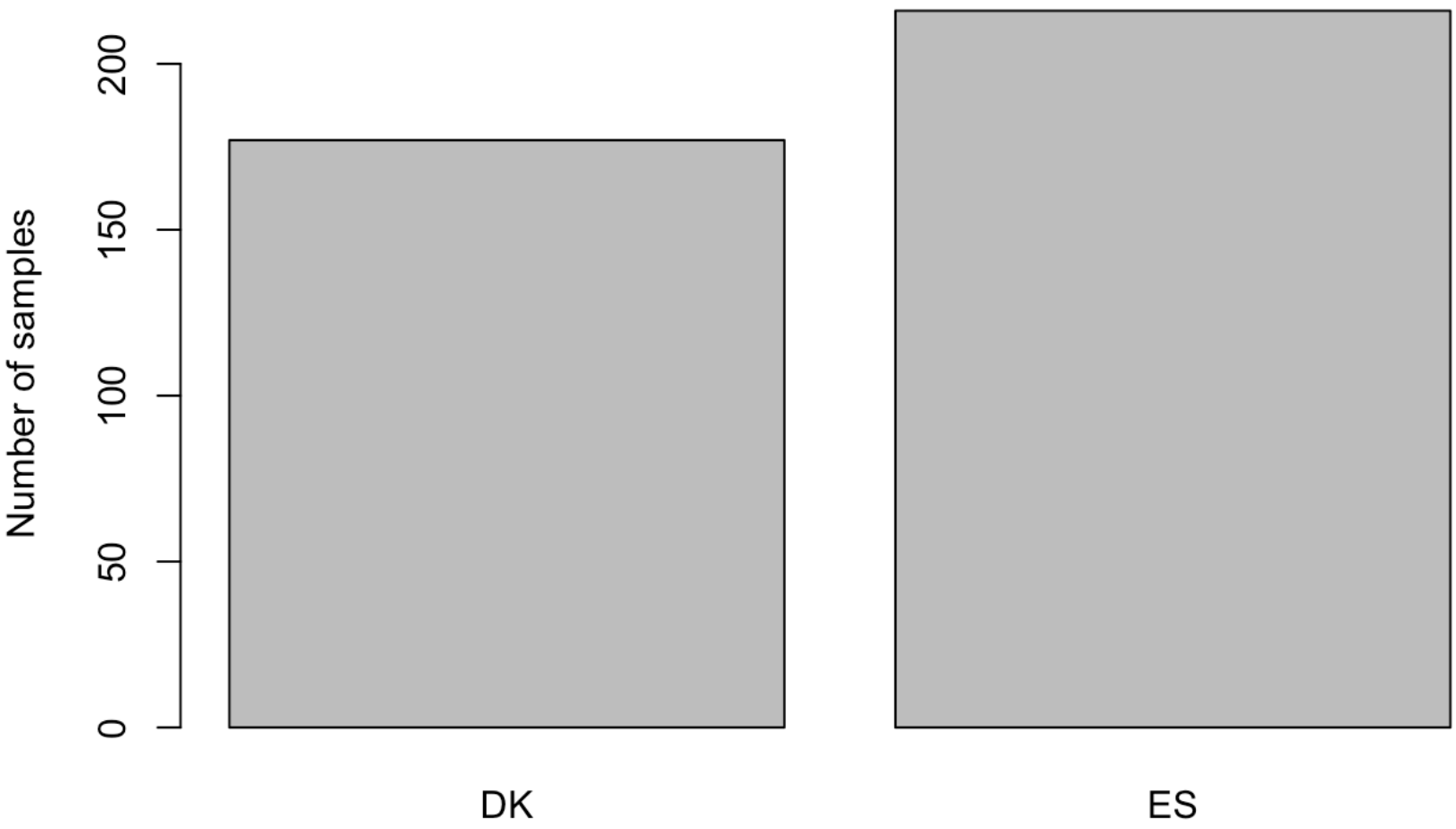
Q3: There are 248 Healthy, 124 UC and 21 CD individuals. This is not a balanced design because of the large differences in the group sizes (e.g. ~11 more Healthy than CD).

Q4:

```
barplot(table(metadata$Gender), ylab="Number of samples")
```




```
barplot(table(metadata$Nationality), ylab="Number of samples")
```



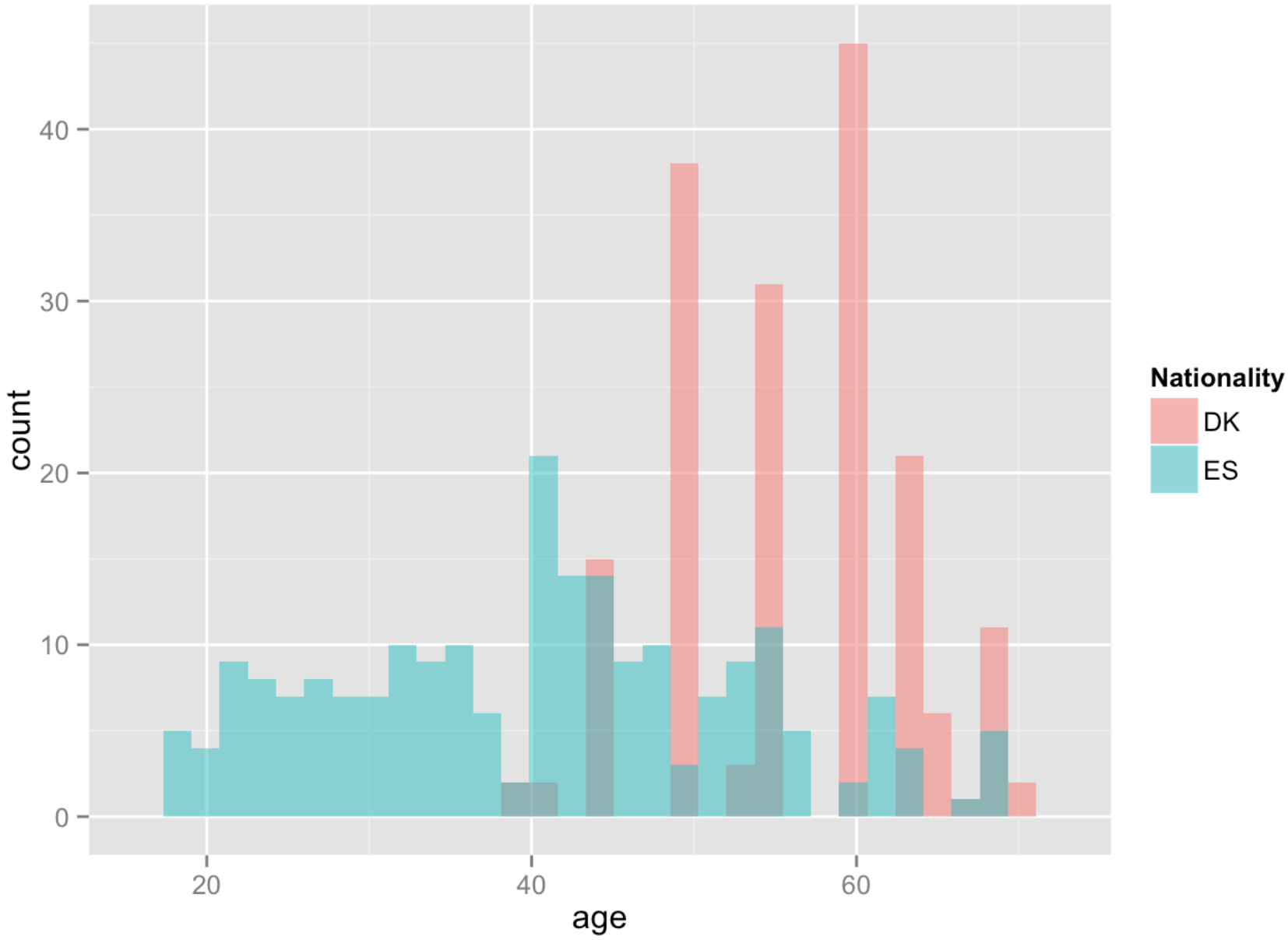
Q5: 177 Danes, 216 Spaniards; 170 men, 223 females

```
table(metadata$Nationality)
table(metadata$Gender)
```

Q6: The distribution of MGS richness looks normal (also called Gaussian) - its shape resembles a bell. The age distribution is flat on the left side and peaks at 50-60. Coloring by nationality reveals that it's the Danish samples that drive that trend. BMI is discussed in the next question.

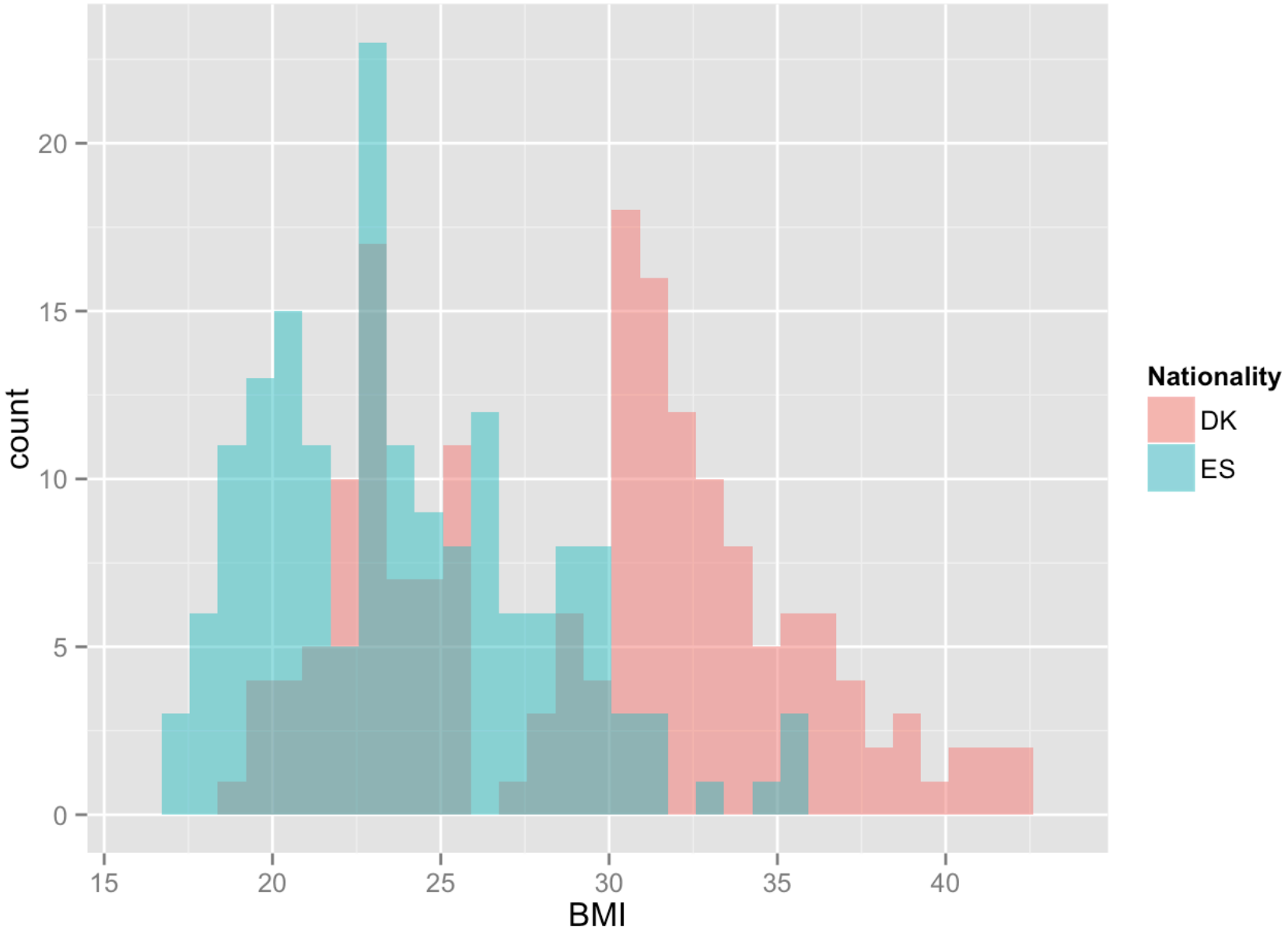
```
ggplot(metadata, aes(x=age,fill=Nationality)) + geom_histogram(position="identity"
, alpha=0.5)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



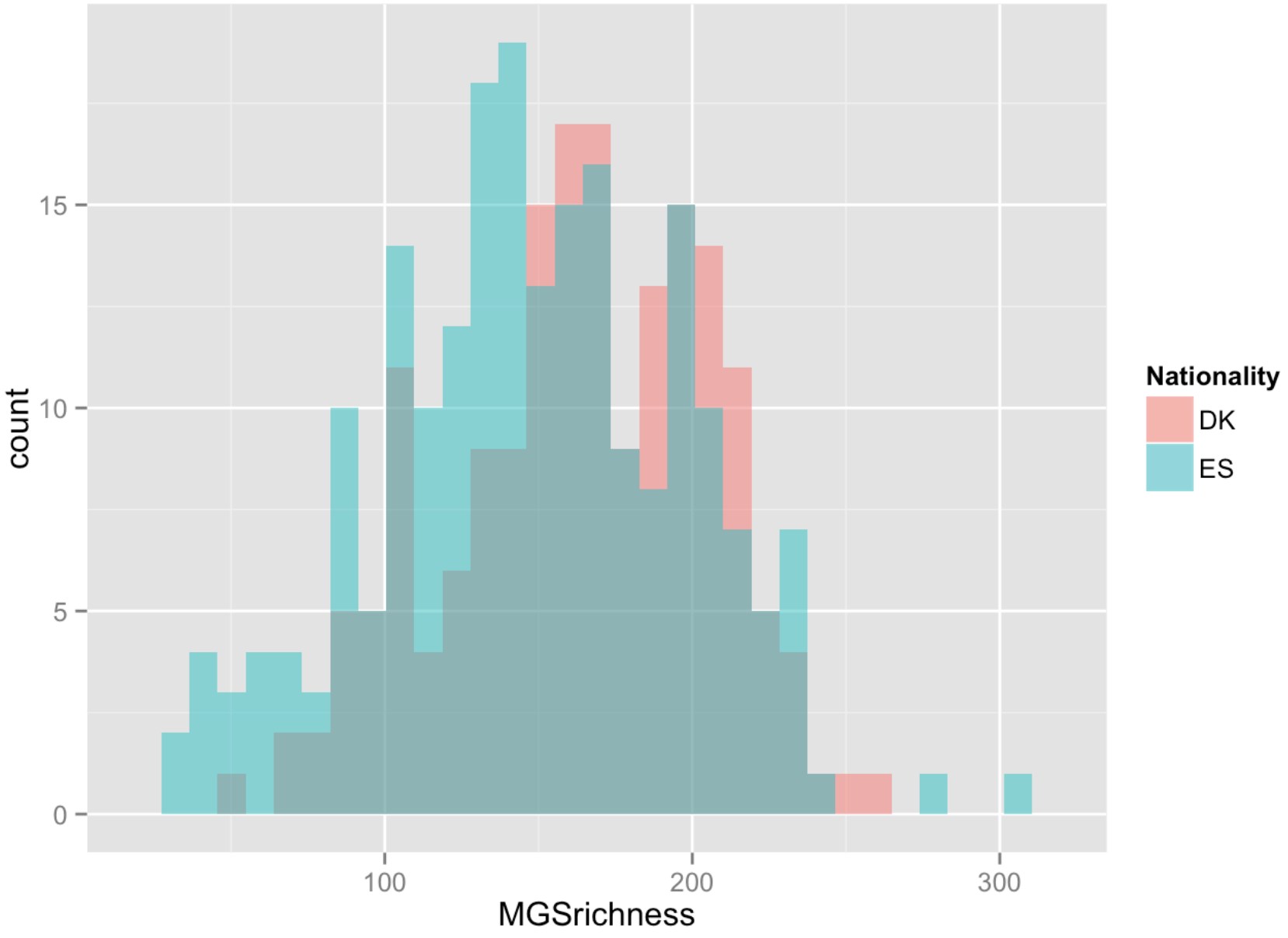

```
ggplot(metadata, aes(x=BMI,fill=Nationality)) + geom_histogram(position="identity"
, alpha=0.5)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```




```
ggplot(metadata, aes(x=MGSrichness,fill=Nationality)) + geom_histogram(position="i
dentity", alpha=0.5)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Q7: The BMI distribution for Danish samples reveals a cluster of samples with BMI<25 (normal weight) and a cluster of samples with BMI>30 (obesity). This distribution doesn't reflect the Danish population but facilitates a comparative study of normal and obese individuals.

Q8: 741 MGS'es, 6640 CAG's

```
table(mgs_cag_geneCount >= 700)
length(grep("MGS", rownames(speciec_matrix)))
length(grep("CAG", rownames(speciec_matrix)))
```

Q9: R syntax can be confusing, as it allows for using output from one function in another function within the same command (line). Here we first selected rows in the column that contain MGS'es (based on 700 gene cut-off) and then turned the matrix into a TRUE/FALSE matrix that indicates if MGS is present/absent in an individual. Finally we count how many TRUE are in each column (in every individual), which gives the count of MGS'es (species) per individual. That's the richness count and it should agree with the MGSrichness column in data frame metadata.

```
species_richness <- colSums(speciec_matrix[mgs_cag_geneCount>=700,] > 0)
species_richness == metadata$MGSrichness
```

Q10: The low richness individuals are primarily young and there are a few high richness individuals that are old. The relationship between the two variables is not too strong.

Q11: There are 2 individuals with no age value (NA) which breaks down the calculation of correlation. You can manually index out theses individuals (see code below), or use the option complete.obs in the cor() function as shown in the example.

```
sample_with_age_value <- !is.na(metadata$age) # exclamation mark means 'not' and c
an be used to reverse the TRUE/FALSE values - try removing it and inspect sample_w
ith_age_value
cor(metadata$MGSrichness[sample_with_age_value], metadata$age[sample_with_age_valu
e])
```

```
## [1] 0.2362206
```

Q12: Correlation between age and species richness is significant (P<0.05) but it's not very strong (r=0.24). Yatsunenko et al looked at the same relationship across wider age range (0-86) in healthy humans and reports an increase of OTU richness in the early life (0-6 years), followed by a rather stable OTU richness trend with increasing age. To make a fair comparison to Yatsunenko et al we would need to only include healthy individuals from our cohort in the analysis. Also, please note that one interesting aspect of Yatsunenko et al is a comparison between different populations - summarized in panel c in Figure 2. Adult Amerindians and Malawians have a higher OTU richness, on average, than US individuals.

Q13: We are only comparing two groups so t-test can be applied instead.

Q14: We observe a significant richness difference between Danish and Spanish individuals. However, we also know that among Spanish individuals we have Crohn's disease and ulcerative colitis patients and both these phenotypes are associated with changes in the gut microbiome structure. So we can suspect that the nationality to richness association is confounded by the disease status in the Spanish group.

Q15: As expected, the CD and UC individuals have a lower species richness than healthy individuals. To draw a conclusion about richness difference between Spanish and Danish individuals we would need to only include healthy Spaniards in the analysis.

Q16: Statistical comparison revealed a significant difference between CD, UC and healthy individuals. The trend is strongest for CD-Healthy and CD-UC comparison, and less so but still significant for UC-Healthy (P<0.05).

Q17: The boxplot shows that abundance of MGS:6 is on average higher in CD than in Healthy individuals. Let's check the species richness as we know that it's decreased in CD patients. The abundance of MGS:6 is negatively correlated to species richness, but this trend is true across the whole cohort, not only CD patients. Based on that observation we can say that richness is better at explaining the variation in MGS:6 abundance than health status.

Q18: We tested MGS:6 abundance association to health status (CD and Healthy) with and without a richness parameter. After adding a richness parameter to the model we observe an increase in pvalue for health status (it remains significant though), confirming our earlier observation that richness is better at explaining MGS:6 abundance changes in the cohort.

Q19: There are 47 MGS'es significantly associated with Crohn's disease after accounting for richness. Since we performed multiple statistical tests for the same variable (health status) we adjusted the resulting pvalues using Benjamini-Hochberg procedure. If we used un-adjusted pvalues, there would be 91 MGS'es significantly different between CD and Healthy individuals.