Radboud University

Wageningen University

Faculty of Science

Department of Animal Sciences

A GWAS about the wingsize of Nasonia Vitripennis

Bachelor Thesis

Author:
Jan Aarts

Supervisors:
Piter Bijma
Shuwen Xia
Johannes Textor
Arjen de Vries

March 2020

1 Abstract

Very little is known about the genetics of wing size in insects. Finding Single Nucleotide Polymorphisms (SNPs) with an effect in wing size can lead to the discovery of genes regulating wing properties in insects. Wing size is an important property because it allows for longer and more efficient flights.

This thesis is a Genome Wide Association study where we will try to find SNPs with a link to aspect ratio, wing to body ratio, in *Nasonia Vitripennis*. The discovery of these SNPs can help in finding candidate genes which can eventually help us find the genes regulating wing size in insects. The study will use data from the *Nasonia Vitripennis*, a model animal in insect biology[1].

The outcome of this study is that there are no single significant SNPs found which indicates that there must be a complex of genes each regulating only a small part in aspect ratio in *Nasonia Vitripennis*.

2 Introduction

Ruined harvests because of insects is a real problem for farmers. Luckily, since the 19th century we have invented insecticides to kill the insects feeding on our crops [2]. These insecticides are usually toxic chemicals designed to kill any insects in the field.

However, these insecticides are also killing non harmful insects and in some cases can be harmful for humans as well. It is clear that this is not a perfect solution and there is an increasing demand for insecticides which do not have so many bad effects on the environment.[3]

This is why there is an increased interest in using parasitiod wasps to combat insects. These wasps lay eggs in insect larvae and therefore decrease the amount of insects in a field. Moreover there are wasps who only target a specific host and that might be a way to target only specific crop eating insects.[4]

If we would be able to replace chemical insecticides with parasitoid wasps then we have a way to keep our crops safe while having minimal side effects in terms of loss of other insects and potentially harmful substances in our crops.[3]

To improve this new 'insecticide' there is a need to learn more about parasitiod wasps. This thesis will focus on finding links between single nucleotide polymorphisms (SNPs) and wingsize in Nasonia Vitripennis. SNPs are single deviations in the genome compared to the reference genome as will be explained in the chapter Preliminaries. The reason why *Nasonina Vitripennis* is chosen is because this is a model animal in the biology and results found in *Nasonia Vitripennis* may apply in other insects as well[1]. Part of the reason *Nasonia Vitripennis* is a model animal is because of its sex determination system called haplodiploidy[5]. Haplodiploidy means that all the males are haploid and hatch from unfertilised eggs and all the females are diploid and hatch from fertilised eggs, which allows us to exploit certain advantages from haploid genetics and study them in more complex systems[6].

This brings us to the research problem: Are there statistically significant SNPs with a link to wing to body ratio, also named aspect ratio, in *Nasonia Vitripennis*? This question will be answered by building a linear mixed model and analysing the P-values for each SNP in a Manhattan plot.

3 Preliminaries

This chapter will provide the knowledge to be able to understand the techniques used in the Methods chapter. The structure is as follows: First, there is a recap of the genetic biology and then an introduction to the concept of genome wide association studies(GWAS) and single nucleotide polymorphisms (SNPs). Lastly, the sequencing method will be explained briefly because it is different than the sequencing method used in human disease research and this technique requires different steps to draw the right conclusions.

3.1 Genomes, chromosomes and alleles

The complete genetic code of an organism is called the genome. The genome is divided in chromosomes, which are really large DNA molecules. For example Humans have 23 different chromosomes and *Nasionia Vitripennis* has five. On these chromosomes the genes are located. Genes are subsections of the chromosome which can be translated to a protein. Not all parts of the chromosome code for a protein and in fact over 98 percent of the total DNA does not code for a protein. In the past these non coding regions where thought to be useless but newer studies are studying its functions in splitting and cutting large chunks of DNA, the production of small RNA's and many other functions[7]. The genes however, are the regions which do code for a protein and these proteins form the instructions or building blocks of life. Proteins regulate processes based on their shape, which is a huge complex with many different folds. Unfolded they are like a really long string. This long string consists of many sub-molecules and these have different properties on their own. Some molecules can bind together, some attract to water and vice versa. This makes it that the protein will fold in many ways and acquires its unique shape. It uses this shape to interact or bind with other molecules and in this way it can regulate processes or build larger molecules. See Figure 1 for an example of a protein fold.

And finally, alleles are just one variant of a gene. Diploid organisms inherit a chromosome from each of its parents and will end up with a pair of every chromosome. In the case of *Nasionia Vitripennis* the females are diploid and every female will inherit one of the copies of her mother and the only copy of her father. Every female has two different copies of each gene and each specific copy of that gene is called an allele.

3.1.1 DNA

DNA or deoxyribonucleic acid is a very important molecule in Biology, it contains the information which defines that species. This information is stored in the four bases of DNA, adenine (A), guanine (G), cytosine (C) and thymine (T). The DNA molecule forms a double helix with the bases in the middle. Figure 2 shows this double helix. Adenine binds with thymine and guanine binds with cytosine, these binds are called base-pairs. Long strands of DNA can form genes which can be translated into proteins.

3.2 What is a GWAS?

GWAS is short for Genome Wide Association Study. This is a study which will look at a lot of genetic variants in many individuals to find a variant linked to a specific phenotypic trait. Usually these genetic variants are so called SNPs, short for single nucleotide polymorphism. This is a variant where a single base pair is different from the reference genome or a position in the genome where more than one base is found in the population. For example in Figure 3 there are five strands of DNA from five individuals which are almost the same but in the third place there is a variant and the base pair can be either C or T in this population. This is what we call a SNP. These can be either causal or non causal, causal SNPs are located on a gene and can really make a different protein due to the different base. Non causal SNPs are not located on a gene and that specific mutation does not encode for a different protein. At first glance it may look like we are only interested in causal SNPs but that is not the case. Translation can cause mutations inside and sometimes outside the translated gene. So the presence of SNPs can be an indication that a gene is located nearby, so that

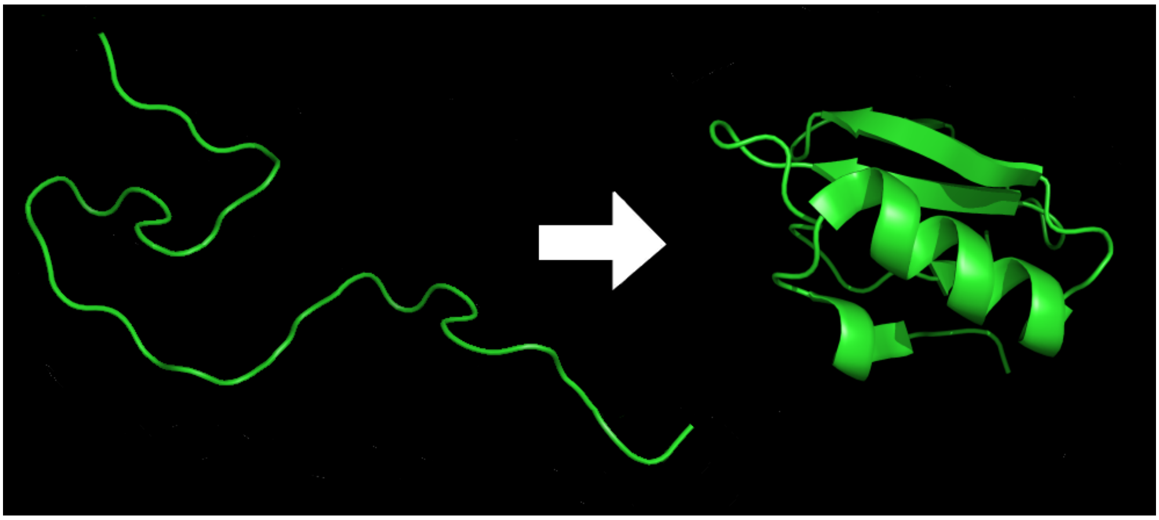

Figure 1: An example of protein folding

if we find a SNP with a significant effect on a phenotypic trait than that makes it likely that a gene with an effect on that phenotypic trait is located near that SNP.

3.3 GBS

GBS is short for Genome by Sequencing and it is a technique which can be used to obtain SNP data. In short this technique will copy and then sample lots of small ∼100 base pairs long DNA strings and will then try to map these together like a giant puzzle. In this way there are several hundreds copies of almost the same string. How many copies exist of the same string is called the read-depth and usually in GBS you want a read-depth of around 100 at minimum. This is because there can be sample errors and with more copies you can recognize or correct these. Compared to the other way of obtaining SNP data with SNP arrays, GBS data will be of less quality. Therefore, it is necessary to apply more strict filters in the quality control step as opposed to a GWAS with SNP arrays. Another difference with SNP arrays is that GBS is way less expensive. In SNP arrays, important known SNPs are selected and placed on a array used to detect the same SNP in other samples. So all SNPs found are manually selected and you need prior knowledge when making these SNP arrays. GBS however does not need prior knowledge about SNPs and can be done on species where little or no SNPs are known such as in our case, *Nasonia Vitripennis*.

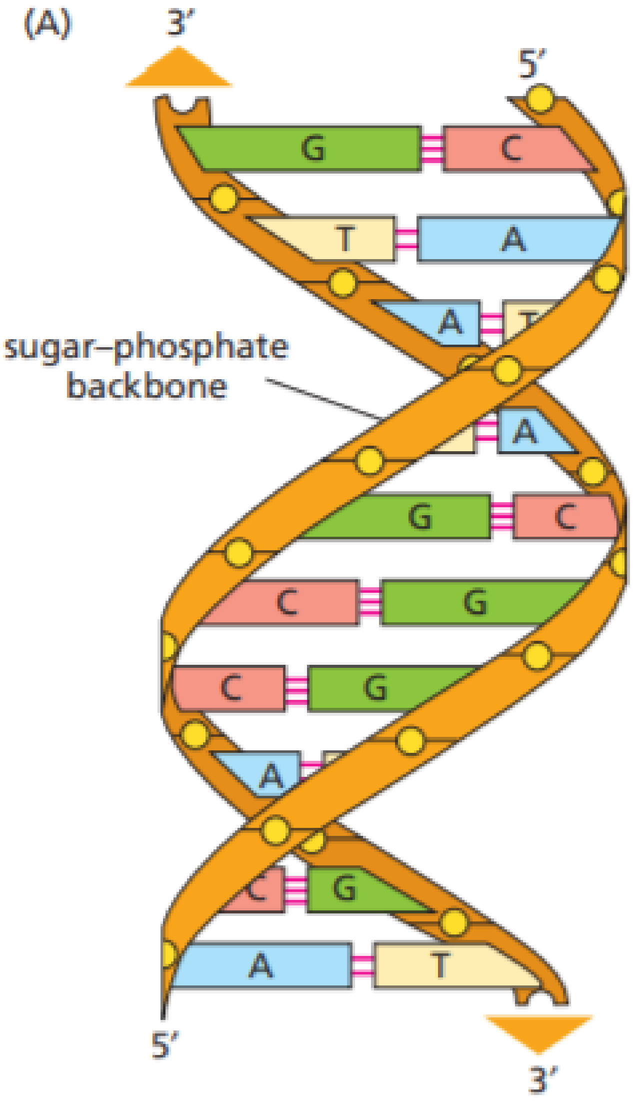

Figure 2: A DNA molecule with the bases in the middle [8].

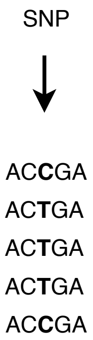

Figure 3: A single nucleotide polymorphism.

4 Methods

This chapter will elaborate on the methods used in this research. The first subsection is about the preparation of the data. The second part is about the model used to calculate significance of the effect each SNP has on the phenotype. The unfiltered dataset and the python scripts used can be found here on gitlab [1].

4.1 Cleaning the Data

Before we can start the analysis the quality of the data needs to be checked. The data consists of the genotype SNP data and the phenotype data. The SNP data is provided by the company BGI using Genotyping by Sequencing (GBS)[9]. As explained in the previous chapter, GBS will result in a lot of unusable data, that is why so many SNPs are filtered out in the quality control phase compared to other genome wide association studies. This is unusual for SNPs obtained via SNP micro arrays, where certain SNPs are preselected based on their known effects or relevance in other studies. The phenotype data is provided by Shuwen Xia [10]. In total there are 1237 individuals and 25287 SNPs. The raw SNP data may contain very rare SNPs with only a few individuals linked to them or SNPs where the minor allele frequency is very low. Those SNPs need to be removed from the data set and I will explain why in the upcoming sections. Most of the quality control steps are done with the free to use software plink [11]. In this thesis I will work with plink 1.9 available on the plink website [2].

4.2 Call Rate

In a GWAS you want to avoid drawing conclusions on SNPs with very few individuals linked to them. SNPs which are present in only a few individuals can look very powerful as seem to be linked to all rare traits those few individuals might have. Those links however are almost always not significant and are very prone to being a false positive. Especially with GBS data where lots of fragments are sampled, then there are a lot of SNPs with a very low call rate. In this thesis I removed all SNPs with a call rate below 90%, which removes all SNPs where less than 90% of the individuals have data for.

4.3 Minor Allele Frequency

We also want to filter on SNPs with a very low frequency on the minor allele. We call the most common allele on a SNP the major allele and the second most common allele the minor allele. In this step I will filter out SNPs with a very low frequency of the minor allele because they could be deleterious and have no statistical power[12].

4.4 Hardy-Weinberg Equilibrium

The Hardy Weinberg equilibrium is a good way to check the quality of the SNPs. It works as follows: First, we assume that the population and thus the SNPs, follow the Hardy Weinberg equilibrium. Then assuming that this equilibrium holds we can detect the SNPs which deviate from this equilibrium and classify them as genotyping errors and remove them from the data set. In this case we use a filter of $0.05/1237 = 0.00004042$ where 0.05 is the significance threshold and 1237 is the number of individuals.

4.5 Population Substructure

When working with genotypes from multiple populations then you should look for population substructures. It could be the case that a normally very rare SNP is much more common in that

[1] https://gitlab.science.ru.nl/yaarts/gwas-on-the-wingsize-of-nasonia-vitripennis

[2] http://zzz.bwh.harvard.edu/plink/contact.shtml

population. This could lead to overestimating the effect of that SNP since the population specific traits could be linked to it. In our case, the wasps are all bred in Wageningen, but since they where raised in two different batches it is still recommended to check for substructures. We can do this by using Principal Component Analysis where the first Principal Component is plotted against the second. This plot can visualise clusters in the data which can be an indication of a population substructure.

4.6 Calculating Inbreeding Factor

A seemingly unusual check we need to perform is the check for inbreeding. Inbreeding is the rate of homozygous alleles in the genome, and this rate increases when siblings mate with each other. This is not a problem on itself for this study but a high inbreeding factor will affect the Hardy Weinberg equilibrium such that a 'good' SNP may be mistakenly classified as a genotyping error. The reason we need to check for inbreeding is because we are working with parasitoid wasps. These wasps are mature as soon as they 'hatch' from their hosts and since siblings hatch from the same host there is a real chance that they will start mating each other immediately. The inbreeding factor(F) can be calculated using this formula:

$F = \frac{(OC–EC)}{(TO–EC)}$

Where OC = observed homozygous count, EC = expected count and TO = total observations. A inbreeding factor higher than zero means there are more homozygous alleles than expected and is an indication of inbreeding. If this would be the case then we should review our filter on the Hardy Weinberg equilibrium because it is most likely too strict.

4.7 Other Quality Control Checks

Some papers will also filter SNPs based on other quality control checks. Since I am not working with human data some of these checks are now irrelevant. In the following sections I will explain which checks are irrelevant and why they are not needed in this study.

4.7.1 Sex Inconsistency Check

A Sex inconsistency check will look at the genotype data and the phenotype data and look if the given sex in the phenotype data is the same as can be predicted from the genotype data. This is irrelevant in our case with data from *Nasonia Vitripennis.* Because of the haplodiploidy of this species, the data contains only females which are diploid so checking for sex based on the genotype is not necessary.

4.7.2 Sample Relatedness

You usually want to know the relatedness between samples since if a certain effect is more common in a family than every unique SNP for that family can explain all these effects. To avoid these false positive effects we want to account for sample relatedness. This step is not necessary in our data since we are working with a linear mixed model instead of a linear model. The linear mixed model will account for sample relatedness in the model itself as random effects. In GEMMA [13] this is done by generating a relatedness matrix.

4.7.3 Batch Effect Analysis

Usually in a GWAS you need to account for batch effects. But in our case the linear mixed model will account for these effects as random effects and no manual checks are needed. Batch effects can include, the temperature in the growing room, the amount of food given, the available space. But depending on the experiment other factors can have an effect.

4.8 Linear models

Linear models are statistical models of the form

$y = mx + b$

In short, the variables x can be used to predict y with a link function. In this thesis we will use both a linear model and a linear mixed model in our analysis. First, we use a linear model to account for our covariates in our phenotypes.

$phenotypes = covariate1 + covariate2 + covariate3 + e$

Then we will use the residuals of this model as the new phenotypes for the linear mixed model. In the section below is explained why this step is needed. Finally, the linear mixed model will fit a model of the form:

$y = W\alpha + x\beta + u + e$

Where y now are our corrected phenotypes. In the section linear mixed model I will explain why we need a linear mixed model here instead of a more simple linear model.

4.8.1 Linear Model

In our data set we have the three covariates namely: batch, plate and date. These covariates can affect the outcome of our study so we need to account for them. In some programs this can be done in the linear mixed model but in this case that was not possible. So in order to solve this we fit a linear model

$y = plate + date + batch + e$

First because the variables are categorical, and none is “better” than the other we should binarize these variables. This process is called one hot encoding and to implement it I used Sklearn [14] for both the one hot encoder and the linear model. The residuals are obtained by subtracting the predicted phenotypes from the original phenotypes. These residuals can now be used as new corrected phenotypes in the linear mixed model GEMMA [13] implements.

4.8.2 Linear Mixed Model

A linear mixed model is an extension of the simpler linear model. Linear mixed models also take into account the fixed effects and random effects. These models are essential when there is dependence in the data set. In a simple linear model, there is the underlying statistical assumption that the variables are independent of each other. A linear mixed model can also work with data that has dependent variables. In a Genome Wide Association study, the data is almost always dependent. In this case the wasps are raised in the same environment. Another example is the family structure or the population substructure, families share a large portion of the same DNA and the same can be the case for large groups of individuals from the same geographic area, which we call population substructure. Since a linear model assumes independence in the variables it would need a data set where we manually corrected for these dependencies. A linear mixed model however can account for these dependencies implicitly and that is why it is used here. When saying implicitly, I actually mean through its random effects. As mentioned earlier, the ”mixed” in linear mixed models means it can work with both fixed and random effects. These random effects are used to explain variance in variables where we do not care for the exact effect but we want to account for it. Random effects include: batch effects, relatedness between samples and time of sampling. A fixed effect can include, accounting in advance for smaller individuals in a colder growing chamber. The random effects are crucial in this analysis because without it we would compare every individuals as if it were completely independent but that is far from the reality. Many wasps share the same variables, growing chamber, possible relatedness, space available per individual, and these variables will most likely have some effect on that individual. So without accounting for the random effects it is impossible to draw thrust-worthy conclusions. To fit a univariate linear mixed model I used GEMMA 0.98.1 [13]. According to the GEMMA tutorial "GEMMA is the software implementing the Genome-wide Efficient Mixed Model Association algorithm [15] for a standard linear mixed model and some of its close relatives for genome-wide association studies (GWAS)". GEMMA's main advantage is that it is much faster than other methods while it is still easy to use with the plink data format. In GEMMA you can fit an univariate linear mixed model of the form:

$y = W\alpha + x\beta + u + e$

Where y are the phenotypes, W a matrix with covariates, α the vector with coefficients, x the vector with marker genotypes, and β the effect of the marker, u is a vector with random effects and e is a vector with errors. To calculate u, the relatedness matrix is used, this matrix makes it possible for the linear mixed model to account for kinship in the individuals.

4.9 Multiple Testing

GEMMA will calculate the effect and the significance of that effect for every SNP in our data set. An accepted threshold for significance is usually 0.05%. However, in our study that is not a fair threshold. We are investigating over 6000 SNPs so we also have that many tests. This brings us to the multiple testing problem; At what p-value can we conclude that the SNP has a genome wide significant effect? In the two subsections below I will discuss both the Bonferroni Correction and the False Discovery Rate.

4.9.1 Bonferroni Correction

Bonferroni correction is a method to solve the problem with multiple testing. After all quality control steps there are 6691 SNPs left, normally we would take a significance constant $\alpha = 0.05$ but with Bonferroni correction this would be 0.05/6691 = 0.00000747272, this means a threshold of 7.5×10^{-6}. Bonferroni correction however is too strict in this case since Bonferroni correction assumes independence between all comparisons [16]. This is obviously not the case since SNPs located close to each other tend to inherit together. So to test significance, Bonferroni correction is not well suited here[17].

4.9.2 False Discovery Rate

Another way to test for significance in multiple testing is the false discovery rate. The false discovery rate controls how many false positives are allowed in the final results. With the Bonferroni correction of $\alpha = 0.05$ then that means that we don't allow 5% of all tests to result in a false positive. In FDR we work with so called q-values. A q-value of 5% means that we won't allow 5% false positive in all significant results, this is of course much less strict than the Bonferroni correction and is more suited for Genome Wide Association Studies especially in this case where we are looking for candidate genes in a largely unannotated genome and any result could be worth looking into.

4.10 Program settings

This short section will list the program settings I used for both Plink[11] and Gemma[13]. Plink settings:

- –geno 0.1 (for Call rate)
- –maf 0.01 (minor allele frequency)
- –hardy 0.00004042 (hardy weinberg equilibrium)

- –het (inbreeding factor)
- –check-sex (sex inconsistency check)
- –genome (sample relatedness)

Gemma settings:

- –bfile (plink binary files)
- -gk 1 (compute relationship matrix)
- -o (output name)
- -lmm 4 (linear mixed model with all 3 tests)
- -k (add relationship matrix to calculation)

Example code to generate relationship matrix:
$./gemma - 0.98.1 - linux - static - -bfile < plinkfiles > -gk1 - orelationshipmatrix$
Example code to run linear mixed model:
$./gemma - 0.98.1 - linux - static - -bfile < plinkfiles > -k < path/to/relationshipmatrix >$
$-lmm4 - o < outputname >$

5 Results

In this chapter I will present the significant SNPs found with the linear mixed model, but before we look at these SNPs we must look at the heritability of wingsize in *Nasonia Vitripennis.* The next section will show the results in a Manhattan plot and in the last sections I will explain some verification steps to strengthen my findings.

5.1 Heritability

First, we need to know how much of the phenotype is actually inherited and not caused by outside effects. This is called heritability and it measures how much of the phenotypic traits are determined in the DNA. Shuwen Xia researched this and her study shows that the heritability of wing size in *Nasonia Vitripennis* is between 22% and 25% [10]. This is very important to know because if the heritability would be very low then there is no point in looking for associations in the DNA because there are likely none. But in this case, we now know that certain traits are inherited and therefore we have a chance at finding the genes responsible for the trait 'aspect ratio' in *Nasonia Vitripennis.*

5.2 Significant SNPs

The results from GEMMA are stored in .assoc.txt files and can be found here[3]. To visualize these, I made a small python script to create a Manhattan plot. Since there is uncertainty about what multiple allele frequency (MAF) is a correct threshold I created plots for MAF 0.05, 0.02, 0.01 and 0.005. The results shown here are with MAF 0.01 and the others will be in the supplementary figures.

Figure 4 shows the Manhattan plot with on the x-axis the chromosomes and the relative position of that SNP on the chromosome. On the y-axis is the -log of the significance of the p_score test. Since there are multiple tests we cannot say that any resulting p-value above 0.05 is significant. Typically, with multiple testing either Bonferroni correction or false discovery rate (FDR) is used. Following Bonferroni correction any SNP with a p-value above 7.5×10^{-6} would be considered significant, see previous section about Bonferroni. As can be seen in the Manhattanplot no SNP comes even close to this threshold because significant SNPs would have a p-value above 6 in the graphs y-axis. We also discussed the false discovery rate where we adjust the p-values into q-values. Using the script "FDR.py" with $/alpha = 0.05$ we can plot the adjusted q-values and see if they reject the hypothesis. None of the resulting q-values rejected the hypothesis and we must conclude that there are no SNPs with a significant effect following these corrections for multiple tests.

5.3 Verification

In this subsection I will elaborate on the verification steps. First we will inspect the quantile-quantile plot (Q-Q plot) and then we will do the analysis again but then on a data set where I manually added an effect. Lastly, I will recap on some quality control steps and explain why I chose certain thresholds.

5.3.1 Q-Q plot

In the Q-Q plot we can see the distribution of the data. In this case we expect an uniform distribution. Analyzing the Q-Q plot is crucial to make sure that there are no confounders here. For example, if one batch would be a bit warmer then that could result in slightly larger animals. Then all genetic differences could be associated with the larger size of these animals whereas in reality they would not have any correlation. Even though I looked at these problems in the quality control sections it is a good practice to look for strange deviations in the Q-Q plot. Figure 5 shows a Q-Q plot of the p-values from the linear mixed model and you can clearly see that it follows the expected uniform distribution. This also strengthens our previous findings that there are no SNPs with a significant

[3] `https://gitlab.science.ru.nl/yaarts/gwas-on-the-wingsize-of-nasonia-vitripennis`

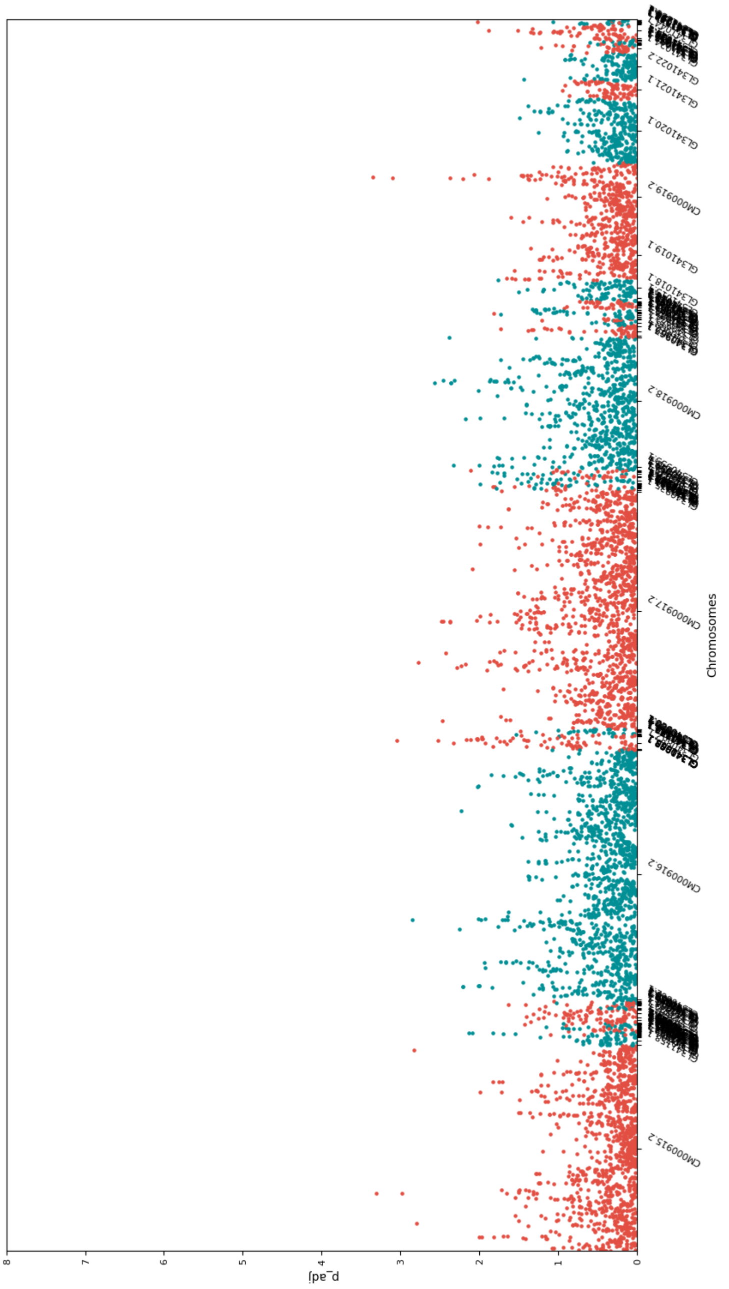

Figure 4: Large Manhattanplot with MAF 0.01

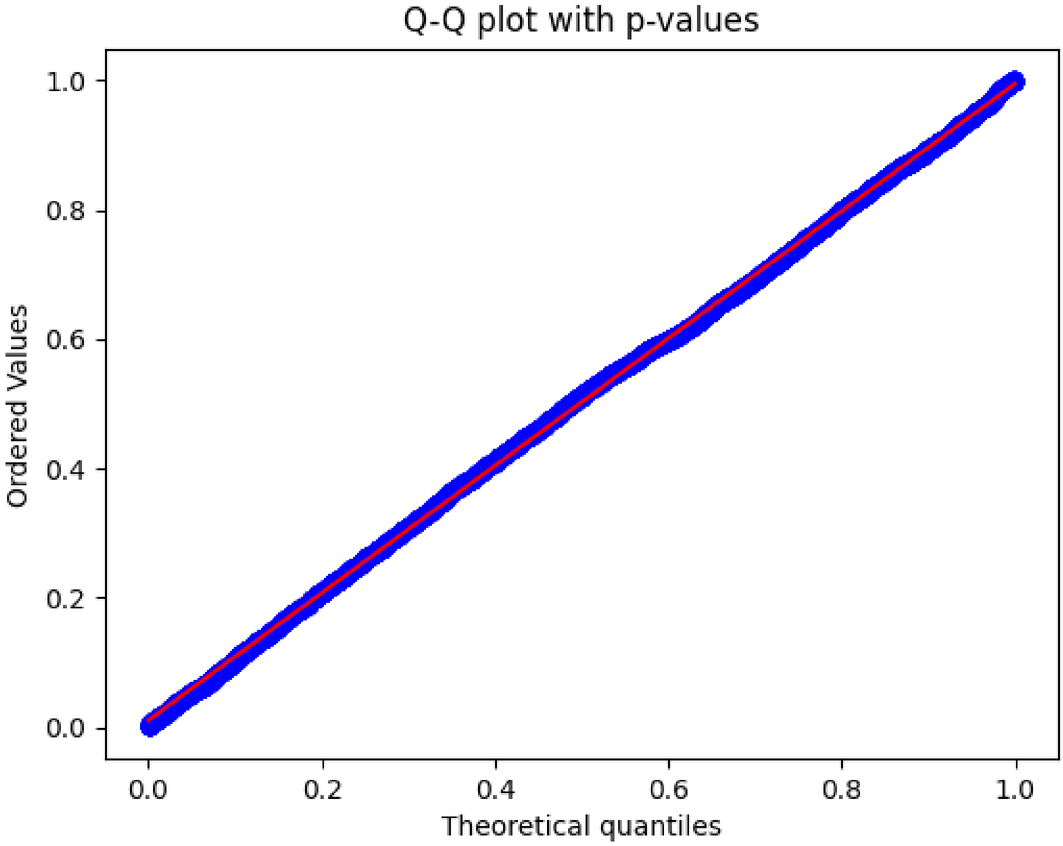

Figure 5: Q-Q plot or probability plot of the resulting p-values

effect on the variance of aspect ratio. Typically, in a GWAS we expect steeper curve near the top to show the effect in the measured phenotype. As can be seen in Figure 6 [18].

5.3.2 Added effect

Since the Manhattan plot shows no significant effect, I wanted to make sure that results are indeed picked up by the model. To test this, I made a data set where I manually added a 0.01 or a 0.02 increase in aspect ratio to every individual with a specific allele. More precisely, I changed the phenotypes of individuals which had the G allele on the 42nd SNP. The variations for this SNP are A A, A G and G G. Each occurrence of G received an 0.01 increase in aspect ratio. My hypothesis is that the SNPs with the added effect although quite small will have a very significant effect in the Manhattan plot. Figure 7 shows the Manhattan plot with the added effect. As can be seen in the Manhattan plot there are quite a few SNPs with a significant effect on aspect ratio. Typically, as in other GWAS there are peaks consisting of a few SNPs which are located close to each other on the chromosome. This is because nearby regions are more likely to inherit together as proven by the linkage disequilibrium. So, although I only added an effect to individuals containing the 42nd SNP there are many more SNPs with a significant effect. Most likely those SNPs are also present in individuals with SNP 42, this can even be the case when those SNPs are not even located on the same chromosome. In conclusion it is very clear that the model can detect SNPs with a variance in the effect of aspect ratio, and that the results shown previously can indeed be trusted.

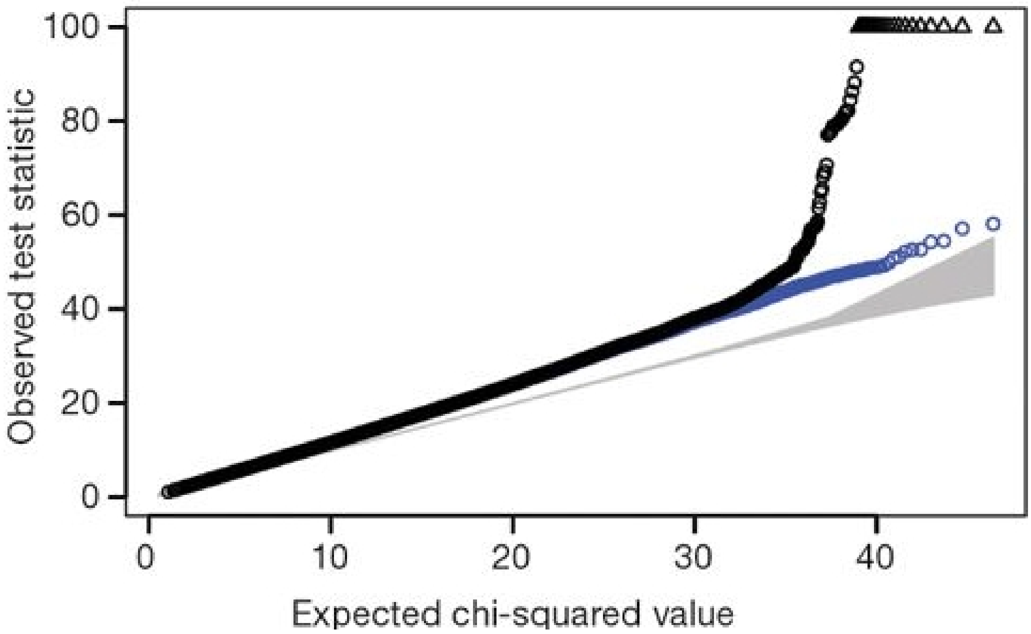

Figure 6: Q-Q plot with significant effect

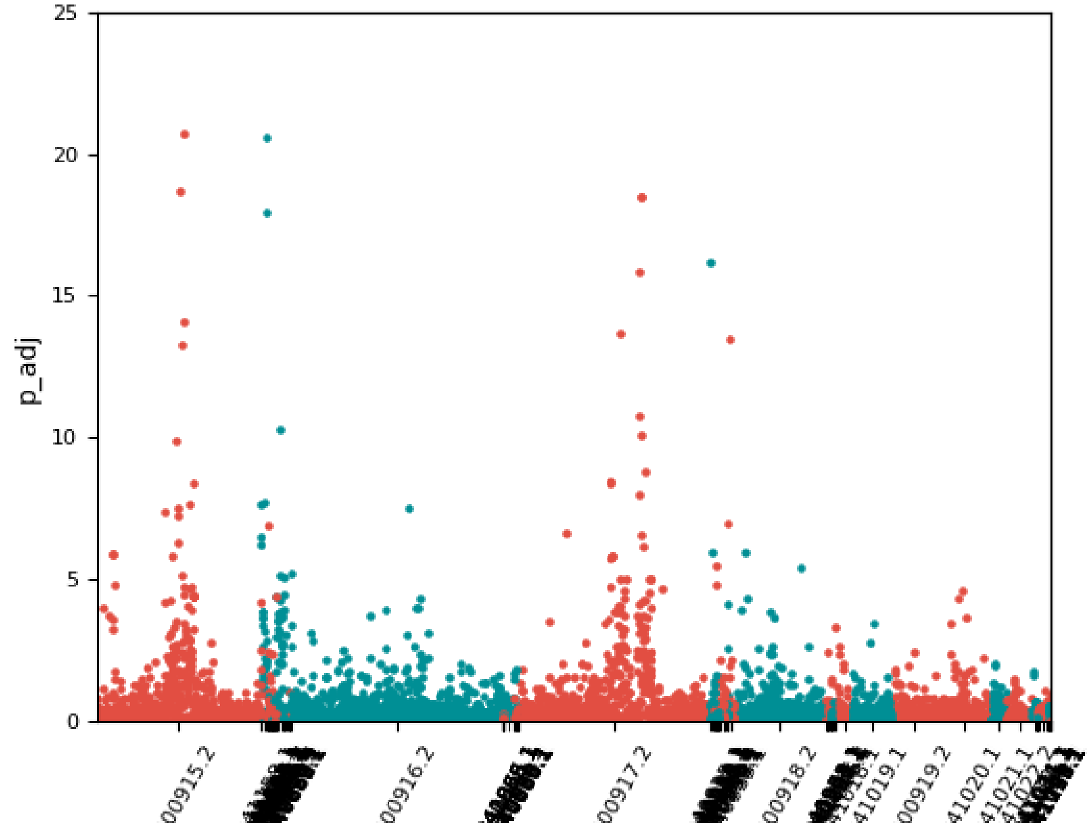

Figure 7: Manhattan plot with an added effect to SNP 42

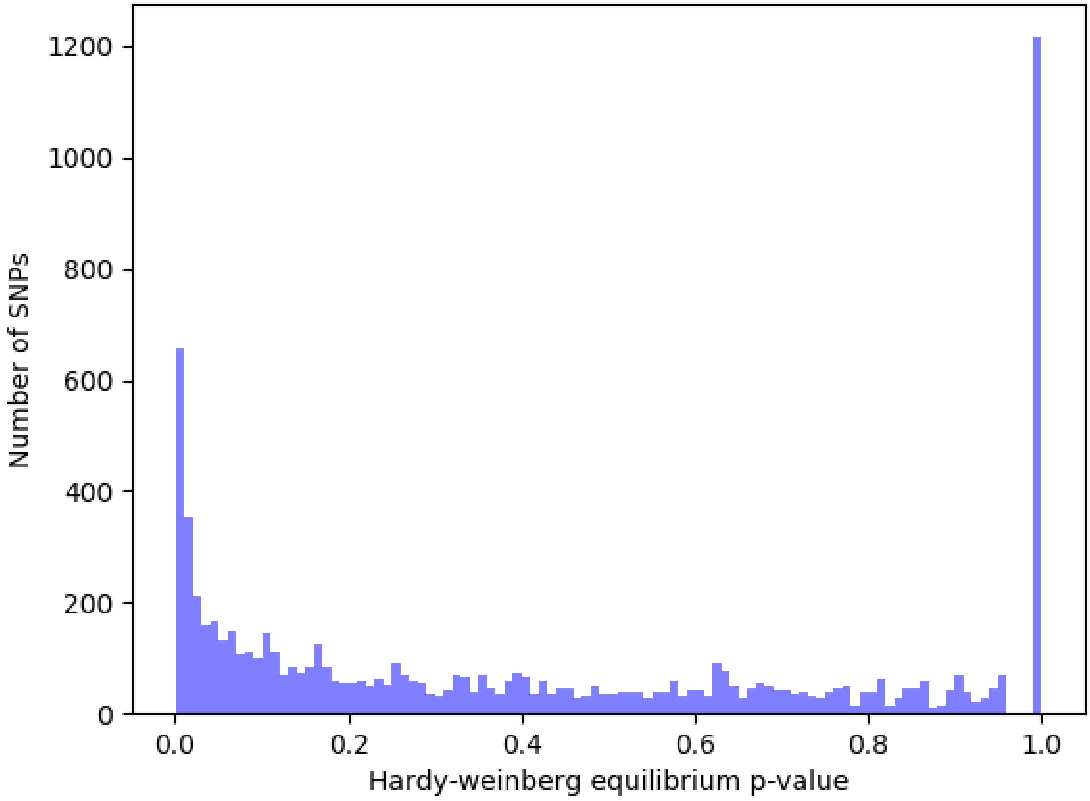

Figure 8: Histogram with p-values of the Hardy Weinberg equilibrium

5.3.3 Analysis of the Hardy Weinberg equilibrium threshold

This subsection will investigate the SNPs and their p-value with respect to the Hardy Weinberg equilibrium. Figure 8 shows a histogram with the number of SNPs and their respective p-value. This graph shows a high spike of SNPs with a very low p-value. These are most likely caused by our sequencing method, Genotyping by Sequencing, and these are removed earlier with our filter of 0.05/1237.

5.3.4 Minor Allele Frequencies

In the methods chapter it was mentioned that I was unsure about the correct threshold to filter for SNPs based on their minor allele frequency. Then I decided to use the filters 0.005, 0.01, 0.02, 0.05 for all steps. In this section we will investigate the histogram of allele frequencies to see which SNPs are most likely sequencing errors and which should be removed. Figure 9 shows a histogram with how many SNPs have a minor allele frequency in a certain range. Here we can see clearly that most of the SNPs have a minor allele frequency of below 0.01, these SNPs are most likely deleterious and are therefore removed [12].

5.3.5 Principal Component Analysis

To detect subgroups in the data we calculated the principal components of the relationships with the relationship matrix. In Figure 10 the first two principal components are plotted against each other. The plot shows no clustering and thus no recognisable subgroups in the data which is good

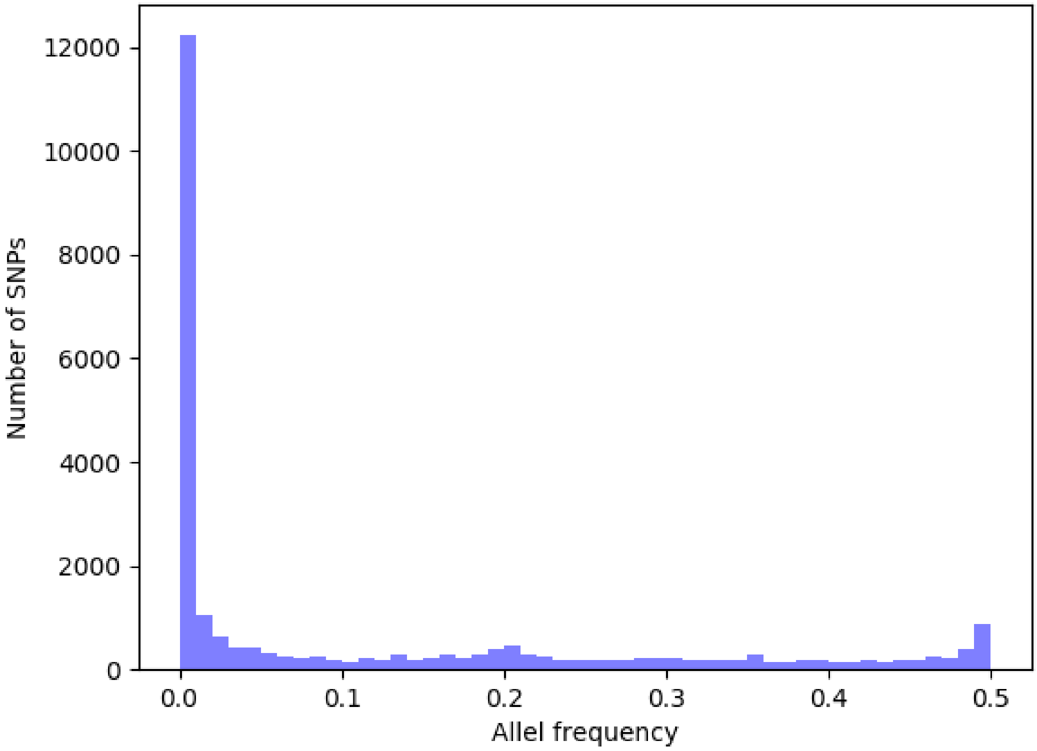

Figure 9: Histogram of allele frequencies.

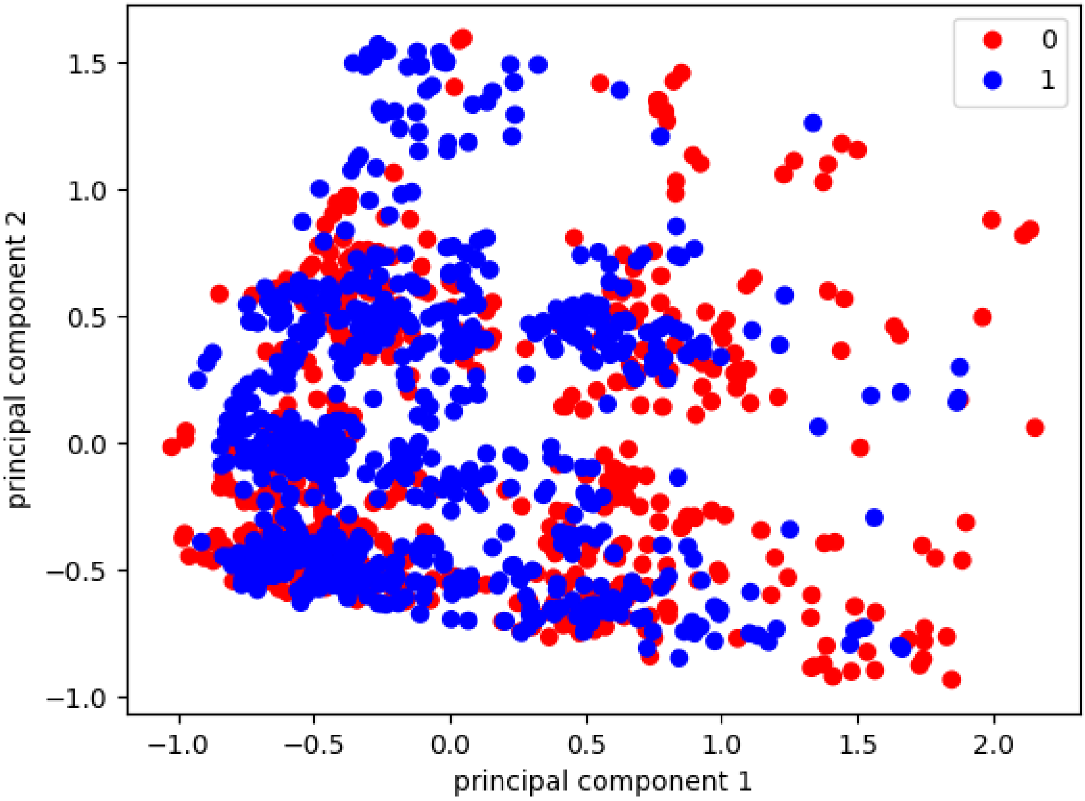

Figure 10: A scatter plot of PC1 and PC2 with the two batches coloured.

since no other methods to account for subgroups are needed. As an extra validation I colored the two batches since they are the most likely to form subgroups for this data, where the individuals are all from the same area. Luckily these batches are not differentiable from each other and we can safely compare individuals between batches.
The final validation check is to make sure that the principal components explain enough of the variance in the data. Figure 11 shows a bar plot with for the first 10 principal components the variance explained. The bar plot shows that the first two principal components explain 45% of the total variance. This is quite high in a genomics study and shows that Figure 10 is trustworthy.

5.3.6 Inbreeding Analysis

The histogram in Figure 12 shows that most individuals have an inbreeding factor of around zero which is the normal case. A few individuals have an inbreeding factor of 0.4 and this is most likely due to chance or Mendelian segregation when the total genome is quite small, as is the case with *Nasonia Vitripennis*, since its genome consists of only five chromosomes. So according to this graph we do not need to be more strict in filtering individuals who deviate from the Hardy Weinberg equilibrium as the inbreeding factor does not indicate a clear inbreeding.

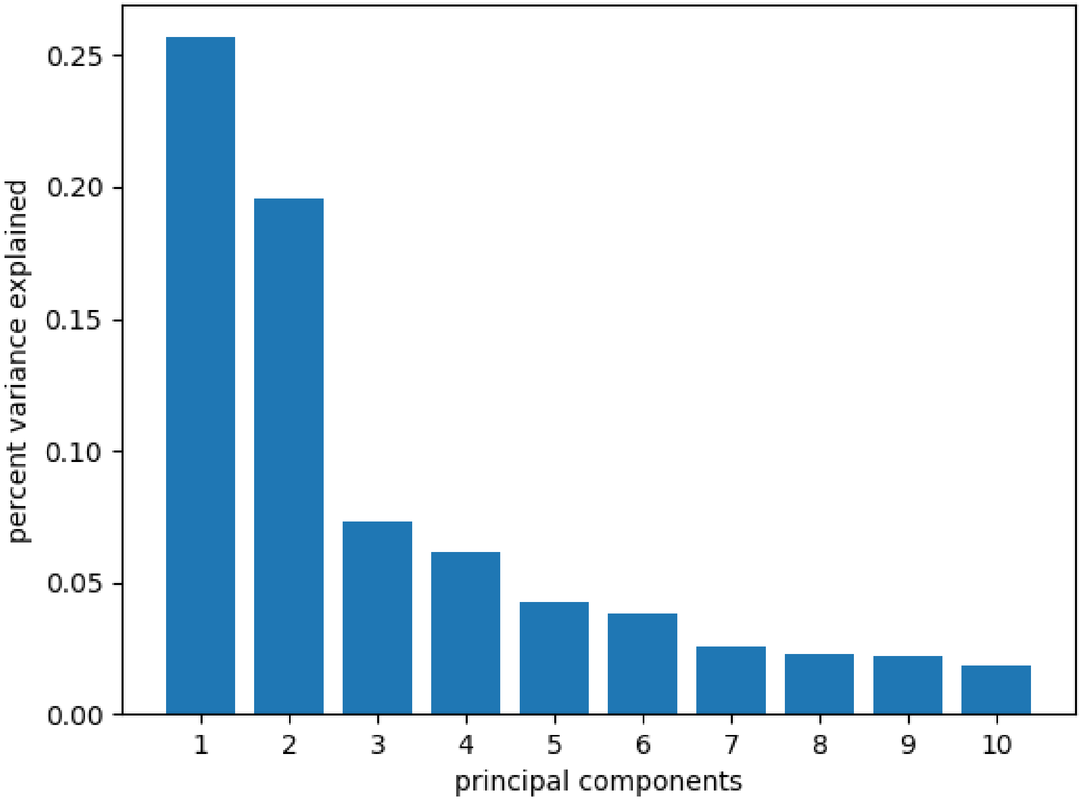

Figure 11: Bar plot with the variance explained ratio for each principal component.

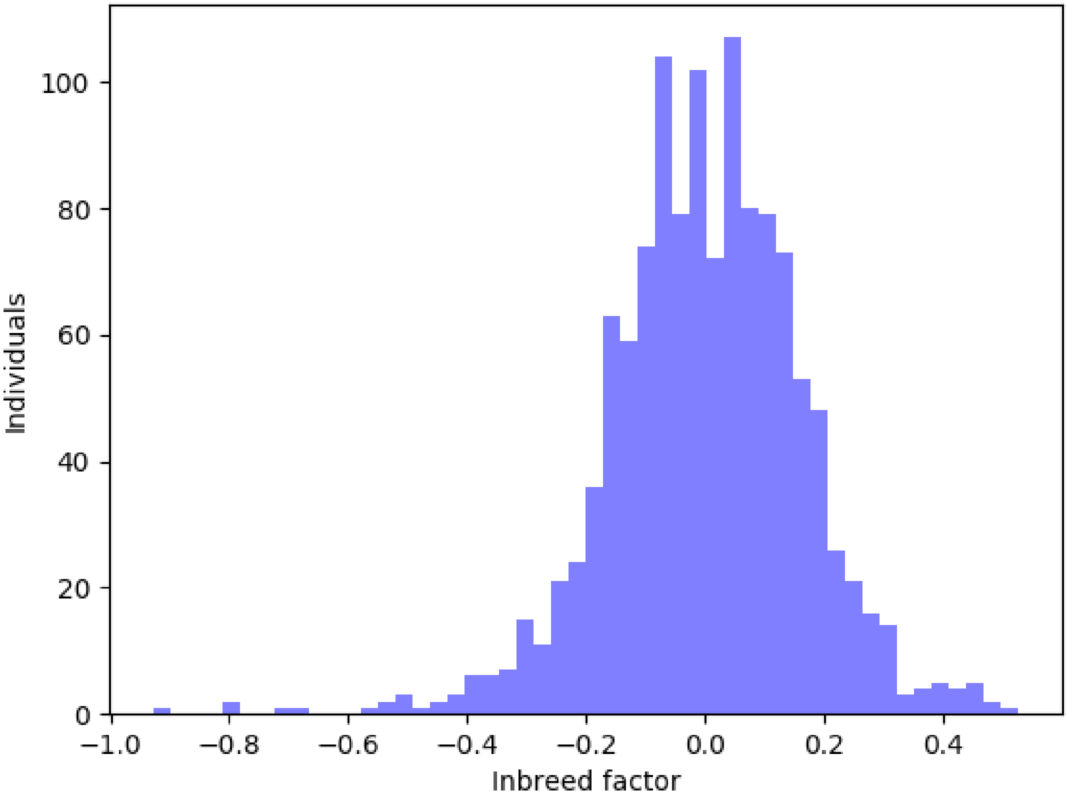

Figure 12: Histogram with animals and inbreeding factor.

6 Conclusion

Even though the linear mixed model did not give any significant SNPs there is still something new to conclude here. In Shuwen' s previous work it was shown that the heritability for aspect ratio is between 22% and 25%[10]. This proves that there must be genes associated with the growth of wings and a link to aspect ratio. But in this study, we could not find any SNPs with a significant link to aspect ratio. Now there are two possibilities:

1. No SNPs are sampled near this hypothetical gene with a link to aspect ratio and therefore it does not show up in the results.
2. There is not a single or a couple genes with a significant effect on the variance in aspect ratio.

Case one is very unlikely because the SNPs cover the whole genome such that there are always multiple SNPs near each position. This means that we can be certain that case two holds and that there must be many genes responsible for aspect ratio each with a very small effect. Similar cases are not unusual in Biology where a certain feature is cause by a complex of genes as is the case with type 2 diabetes [19]. Unfortunately, this makes it impossible to genetically modify individuals with a hypothetical gene which would improve aspect ratio. Therefore, in order to breed individuals with larger wings used in biological control we should rely on family structures as this thesis proves that there no single or few genes associated with aspect ratio exist.

References

[1] Christopher A Desjardins et al. "The genetic basis of interspecies host preference differences in the model parasitoid Nasonia". In: *Heredity* 104.3 (2010), pp. 270–277.

[2] Kurt Brand and Walter Bausch. "Über Verbindungen der Tetraaryl-butanreihe. 10. Mitteilung. Über die Reduktion organischer Halogenverbindungen und Über Verbindungen der Tetraaryl-butanreihe". In: *Journal für Praktische Chemie* 127.1 (1930), pp. 219–239.

[3] Peter Fantke, Rainer Friedrich, and Olivier Jolliet. "Health impact and damage cost assessment of pesticides in Europe". In: *Environment international* 49 (2012), pp. 9–17.

[4] Michael R Strand and John J Obrycki. "Host specificity of insect parasitoids and predators". In: *BioScience* 46.6 (1996), pp. 422–429.

[5] Leo W Beukeboom, Albert Kamping, and Louis van de Zande. "Sex determination in the haplodiploid wasp Nasonia vitripennis (Hymenoptera: Chalcidoidea): a critical consideration of models and evidence". In: *Seminars in cell & developmental biology*. Vol. 18. 3. Elsevier. 2007, pp. 371–378.

[6] *Nasionia research*. URL: `https://www.sas.rochester.edu/bio/labs/WerrenLab/WerrenLab-NasoniaResearch.html`.

[7] Gerald Litwack. *Human biochemistry*. Academic Press, 2017.

[8] Marketa J Zvelebil and Jeremy O Baum. *Understanding bioinformatics*. Garland Science, 2007.

[9] *Genome sequencing Biotechnology*. URL: `https://www.bgi.com/us/`.

[10] Shuwen Xia. "Exploring the potential of genetic improvement of insects". In: ().

[11] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[12] Stephen Turner et al. "Quality control procedures for genome-wide association studies". In: *Current protocols in human genetics* 68.1 (2011), pp. 1–19.

[13] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nature genetics* 44.7 (2012), pp. 821–824.

[14] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[15] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nature genetics* 44.7 (2012), pp. 821–824.

[16] RC Johnson et al. "Accounting for multiple comparisons in a genome-wide association study (GWAS) BMC Genomics. 2010; 11: 724. doi: 10.1186". In: ().

[17] R Bedre. *Bioinformatics data analysis and visualization toolkit*. URL: `https://github.com/reneshbedre/bioinfokit`.

[18] Jeff Barrett. *How to read a genome wide association study*. URL: `http://genomesunzipped.org/2010/07/how-to-read-a-genome-wide-association-study.php`.

[19] Liana K. Billings and Jose C. Florez. *The genetics of type 2 diabetes: what have we learned from GWAS?*

7 Supplementary figures

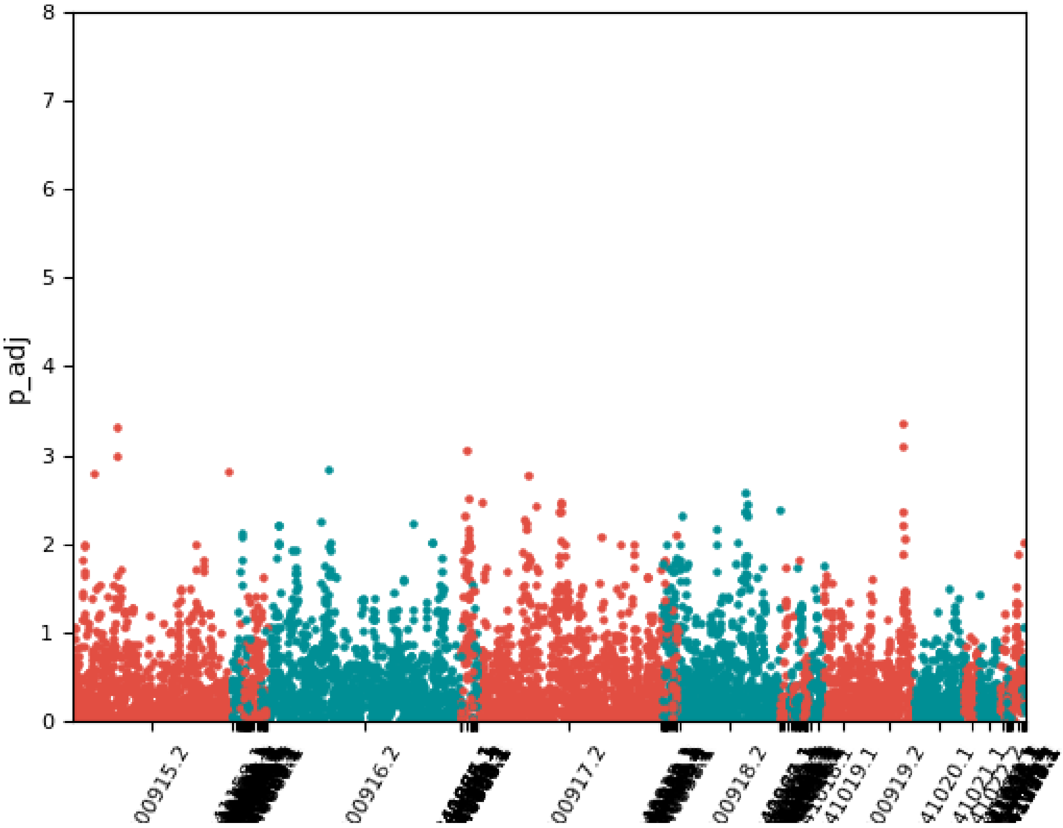

Figure 13: Manhattan plot with MAF 0.005

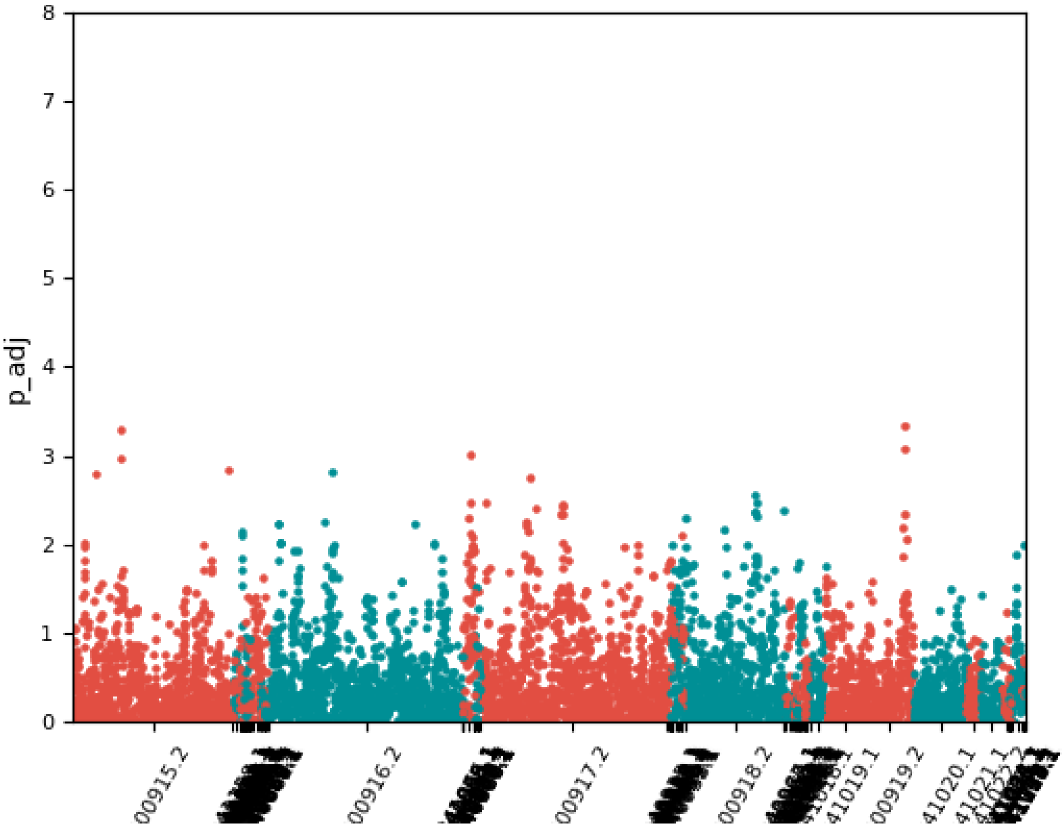

Figure 14: Manhattan plot with MAF 0.02.

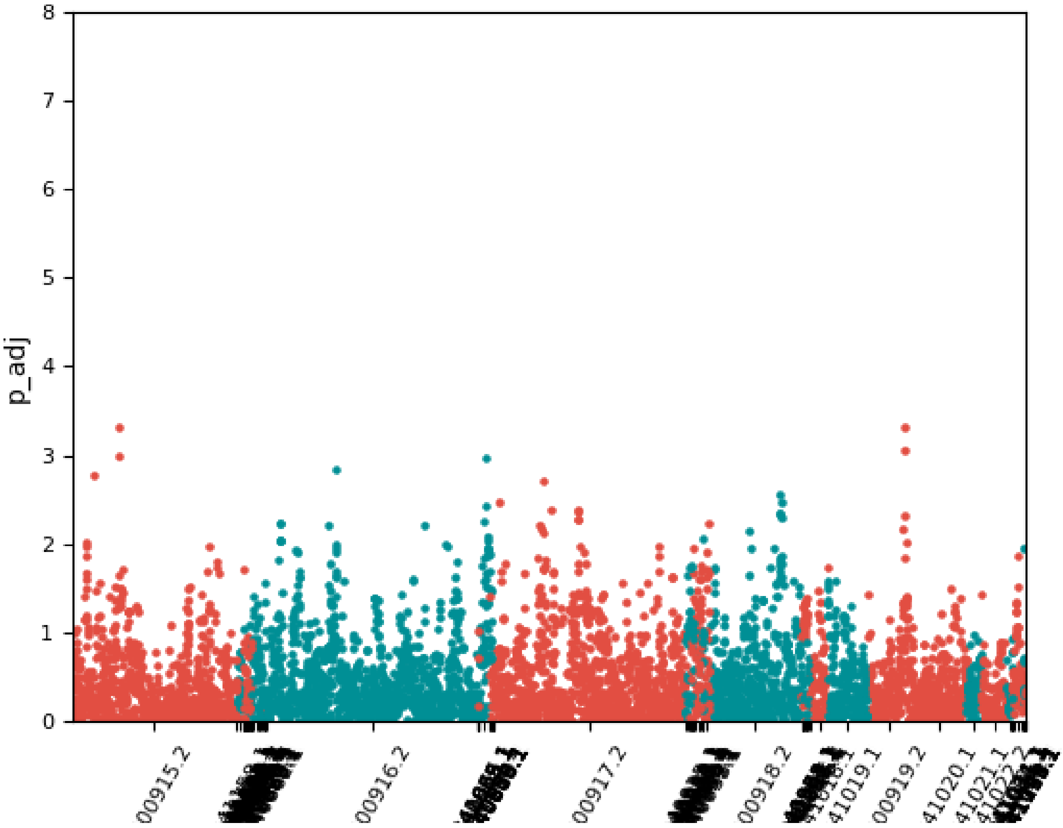

Figure 15: Manhattan plot with MAF 0.05

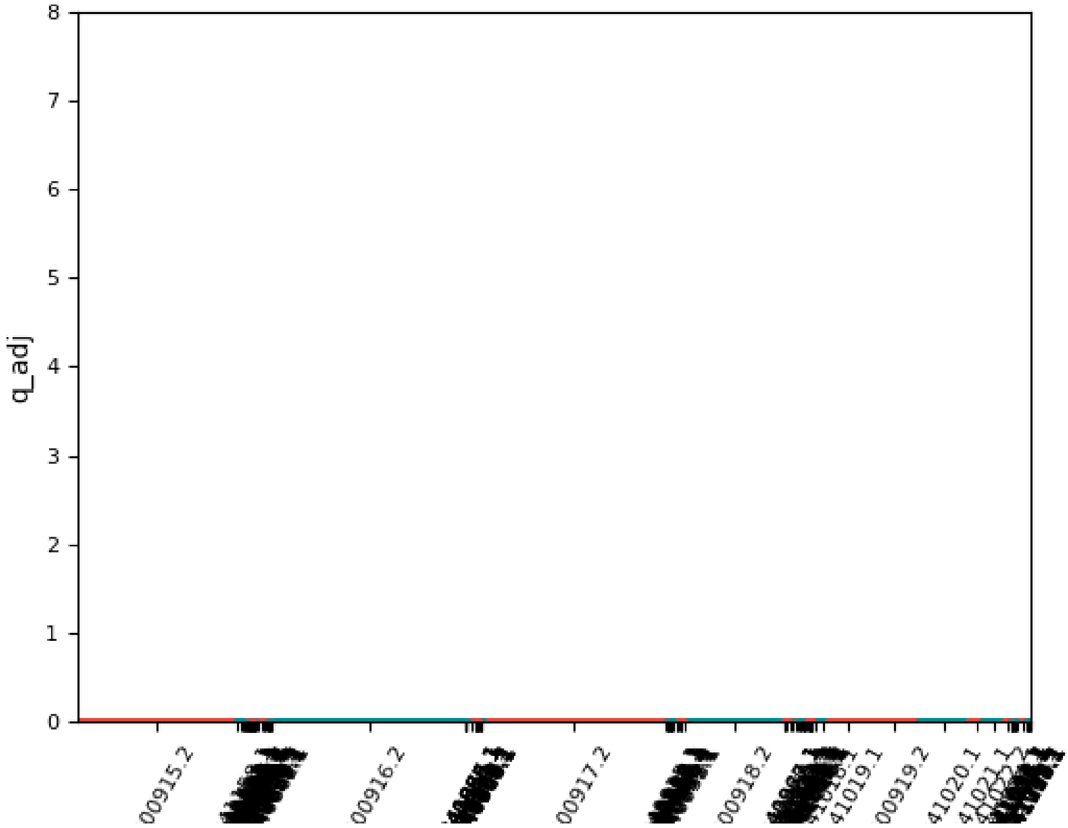

Figure 16: Manhattan plot with q-values and MAF 0.01