SNP MARKER DETECTION WITHIN A HIGHLY INBRED *ASOBARA TABIDA* STRAIN

USING SNP MARKERS FOR LINKAGE GROUPING IN A *DE NOVO* ASSEMBLED TMS GENOME

JITSKE VAN DER LAAN

850121493130

Minor thesis Genetics: GEN-80424

Rijksuniversiteit Groningen

Centre for Evolutionairy and Ecological Studies (CEES)

Nijenborgh 7, Groningen

MARCH 2013

Supervisors/Examiners:

DR. PIETER NEERINCX (UMCG)

DR. EVELINE C. VERHULST (RuG)

DR. LOUIS VAN DE ZANDE (RuG)

PROF. DR. BAS J. ZWAAN (WUR)

WAGENINGEN UNIVERSITY
WAGENINGEN UR

university of groningen

Abstract

Next Generation Sequencing (NGS) technologies have contributed to the rapid increase of sequenced genomes. Since sequence data are produced fast, easy and relatively cheap, challenges are shifted towards genome assembly. Most applications for genome construction make use of a reference genome to align the millions of reads that NGS produces. However, the absence of reference data in *de novo* genome assembly projects makes successful alignment much more complicated. Recently, the genome of the parasitoid wasp *Asobara tabida* has been sequenced for the first time. Assembly resulted in over 600000 scaffolds, possibly because the DNA contains numerous repeats and is divided over many chromosomes. A method for reducing this number of scaffolds is marker-assisted linkage grouping. The purpose of this thesis concerns the efficiency of SNP marker detection in the *de novo* assembled *Asobara* genome and the usability of these SNP makers in the construction of linkage groups. A selection strategy was designed for detection of SNP markers in the assembled genome. False positive SNP detection was reduced by parameter settings that indicate PCR-, sequence-, or assembly errors. The remaining SNP markers were genotyped in a diploid hybrid female wasp and her haploid sons, that all possess a recombined maternal genome. The data were used to order the corresponding scaffolds in linkage groups. Analysis of 10 SNPs showed us that 50% could be assigned to a single linkage group, indicating one large chromosome, whereas the rest of the SNPs were ungrouped. Despite the grouped scaffolds covered only 1.61 ‰ of the total genome length, the evidence for linkage was strong (LOD=8.0), making a small step forward in completion of the *Asobara tabida* genome. However, for a proper prediction of the chromosome number in this species more SNP markers are needed. Although we succeeded in finding SNP markers (>500000) in a highly inbred strain *in silico*, molecular detection showed limited heterozygosity of the SNP markers. For a higher efficiency of heterozygous SNP marker detection it is recommended to include genomic SNP marker information of outbred strains.

Introduction

Hymenoptera are characterized by their haplodiploidy. The parasitoid wasp *Asobara tabida,* one of the Hymenoptera members, parasitizes *Drosophila* larvae to lay eggs. In haplodiploid species both haploid and diploid individuals are present. Egg fertilization result in development of diploid female wasps and unfertilized eggs become haploid males, which is characteristic for an arrhenotokous sex determining system. Sex determination systems in *Asobara* species were outlined by several studies (Beukeboom et al. 2001; Van Wilgenburg et al. 2006). Recently, it was excluded that sex determination in *A. tabida* was controlled by a Complementary Sex Determination (CSD) system (Ma et al. 2013). In a non-CSD system other factors than the sex determination locus affect sex determination. In literature, several of these factors are described in a model, concerning fertilization, genetic balance, maternal effect, or genomic imprinting (Beukeboom et al. 2007).

One of the goals in the study to non-CSD systems is the functional conservation of sex determination genes (Geuverink 2011). To ease and support the detection of these sex determination genes in *A. tabida*, the genome of a highly inbred strain was sequenced recently. As a consequence of repetitive sequences, assembly resulted in numerous scaffolds (>600000) with sequence lengths varying from 80 to 200000 bp. The total *A. tabida* genome length was estimated at 300 Mb. Although a karyotype of this genome is lagging behind, the chromosome number *n* was expected to be 16-18 (Kraaijeveld et al. 1999). Grouping the scaffolds on their linkage will create a better view of chromosomal arrangements. Linkage grouping concerns the extent in which certain chromosomal locations -e.g. genes, exons, and other sequences of interest- are linked and inherited together. These chromosomal locations are detected by genetic markers. Examples of frequently used genetic markers are microsatellites, Amplified Fragment Length Polymorphisms (AFLP's), and Single Nucleotide Polymorphisms (SNPs) (White et al. 2007).

During this thesis, we have searched for SNP markers on scaffolds of the *Asobara tabida* genome assembly. SNP markers were preferred because they can be genotyped efficiently and they are spread genome-wide (Morin et al. 2004). We have adapted an *in silico* SNP caller suitable for detection in a *de novo* genome (SUPPLEMENTARY FILES B1). The SNP-caller was deduced from the original SNP calling pipeline of the University Medical Centre in Groningen, which was based on algorithms provided by the Genome Analysis Toolkit (GATK) (DePristo et al. 2011). The GATK pipeline was adapted for running our local assembly database. The output from GATK was examined, from which a strategy was developed to select for SNPs reliable for linkage grouping.

The haplodiploid genome of *A. tabida* is a suitable model for studying recombination events. As only heterozygous SNPs can provide information about recombination events, the DNA of both inbred and

hybrid strains was screened to maximize the number of heterozygous polymorphisms. Based on this genotyping, we tried to give a linkage-group based approach for the total chromosome number and genome length. For this approach we have examined the success rate of SNP calling in a highly inbred strain and criticized the accuracy of using these SNPs in linkage grouping.

Materials and Methods

1. Bioinformatic procedures

For the genome assembly, SNP-calling, SNP selection and linkage mapping several algorithms were used. The first two parts were accomplished by external sources by which the data was provided for this project. Information about the assembly- and SNP calling procedures is provided in SUPPLEMENTARY FILES B1. The predicted SNPs were shown in a 'VCF-file', reporting the quality values (SUPPLEMENTARY FILES B2). The sequence assembly was visualized by Tablet, an alignment viewer designed for Next Generation Sequence data (Milne et al. 2010).

SNP selection

Variant Call Format (VCF) is a standardized format used to store frequently occurring structural sequence variants, including SNPs. For the analysis of this format we used VCFtools, a software suite that has an application programming interface for Perl and Python (Danecek et al. 2011). SNP selection was performed in two rounds; in the first selection round filtration was based on estimated information and default settings. The quality values used for final SNP selection were based on the estimated coverage (~40x) and length (~300 Mb) of the *A. tabida* genome assembly. Parameters were divided in fixed and variable parameters. The fixed parameters filtered the bad quality calls, including Base Call Quality (>30), Genotype, and Mapping Quality (≥50%). As we did *de novo* SNP calling, the values of these parameters were lower compared to the minimal quality values in reference SNP calling. Therefore it was chosen to fix these values in order to exclude the most bad quality calls. The variable filters were chosen to reduce false positive SNPs occurred by both PCR and sequence errors and the existence of duplicated fragments. The conditions of SNPs were focused at the coverage (30-60x), the reference/alternative nucleotide ratio (1:1), and the homopolymer run (≤2) (SUPPLEMENTARY FILES C1). The output was stored in a separate VCF-file. Since a reference database was absent we did not have any genotypic information. With exception on the 'Genotype' parameter (explained in results), the data concerning genotype information was removed from the file.

The second selection round was limited to SNPs located at the three largest scaffolds. All cut-off values were adapted, which was based on the statistical output of the GATK SNP calling pipeline (SUPPLEMENTARY FILES B3). In these settings the total SNP coverage was chanced from a range of 30-60x to a range 30-110x. To control if the largest scaffolds were assembled correctly, we exclusively have focused on SNPs located at the scaffold ends. The SNP number decreased dramatically when the reference/alternative nucleotide ratio filter was used. To promote SNP detection at the scaffold ends

it was chosen to exclude this ratio filter. The found SNPs were added to the final SNP collection of the first selection round.

Linkage mapping

Linkage analysis was performed with JoinMap® v3.0 linkage mapping software (Van Oijen and Voorrips 2001). We evaluated 10 heterozygous SNPs for a population size *n*= 50, and configured the population type as haploid (HAP). Linkage map calculations were done at threshold of fit set to ≤5.0 with LOD scores >1.0 and a recombination frequency <0.4. A final linkage map was constructed with SNP linkage of LOD scores ≥ 3.0.

Statistics

A diversity of statistical data was acquired from the alignment. Picard is a Java-based utility for extracting information from SAM-files, including cumulative coverage and quality statistics of the *de novo* genome assembly. Furthermore, statistics of the SNP selection procedure were performed with the VCF-stats tool in the Perl API module of VCFtools (Danecek et al. 2011).

2. Molecular procedures

The filtered SNPs were tested for heterozygosity in DNA of diploid females. Heterozygous SNPs were identified in 50 of their haploid male offspring. To study the consequences of inbreeding on SNP heterozygosity, diploid females of both a hybrid HKxTMS strain and an inbred TMS strain were analysed (SUPPLEMENTARY FILES A1 and A2). DNA was extracted according to a high-salt DNA extraction protocol (Aljanabi and Martinez 1997) (SUPPLEMENTARY FILES F1) and diluted in 50 µl sterilized Milli-Q water.

Primer design and PCR

In preparation of SNP genotyping, primers were designed in Perl-Primer (Marshall 2004) with an average amplicon size of 370 bp. All primers were aligned to the original *A. tabida* 'reference' genome[1] by running local BLASTn of NCBI (version 2.2.27). Settings were adjusted to short sequence searches. Primers were discarded if either the total primer length annealed more than 3x to the reference genome without mismatches, or more than 5x if number mismatches ≤3. An overview of the

[1] The reference genome mentioned in this study is not an official database reference, but our first version of the *de novo* genome assembly of *A. tabida*

used primers is shown in SUPPLEMENTARY FILES D1. The primers were diluted to 100 µM in nuclease-free water to create a stock solution.

PCR was performed according to a standard protocol (SUPPLEMENTARY FILES F2). The regular PCR programme started with 3 min denaturation (94°C), followed by 40 cycles of 15 sec at 94 °C, 30 sec at 57 °C, and 30 sec at 72 °C, and finished with 7 min at 72 °C. However, some primers worked optimal for different annealing temperatures as indicated in SUPPLEMENTARY FILES D1. All PCR products were tested on a 1% Agarose gel, containing 1x TAE buffer and 10µg/µl EtBr. The Gene ruler 100 bp DNA ladder (Fermentas Life Sciences) was used for detection of DNA fragment sizes.

SNP Genotyping

DNA originated from F1 females was sequenced from both 5' and 3' directions. Because of economic considerations and efficiency, either the forward or reverse primer was selected and used to sequence the F2 male generation (SUPPLEMENTARY FILES D1: Table D1-3). Purification was done with ExoSAP-IT (Fermentas Life Sciences) for PCR products and Sephadex (Amersham Biosciences) was used for sequence product purification. After purification, the DNA was dissolved in 10 µl of HiDi-formamide (Applied Biosystems) and the final sequence product was sequenced with an Illumina ABI3730 sequencer (SUPPLEMENTARY FILES F3). The SNP genotyping data was analysed with Chromas and BioEdit software (Hall 1999).

Results

504100 SNPs were *in silico* found in our reference genome. These SNPs were distributed over 69868 scaffolds, which is approximately 11% of the total number of scaffolds. In the raw SNP output it was seen that not all SNPs were reliable for analysis. Selection for reliability was started with the design of a cut-off strategy. This strategy was divided in two phases, from which the first phase concerned rough selection. In a later stage the cut-off values were determined as such, that we found specific SNPs relatively close to the ends of the largest scaffolds.

SNP selection - Fixed parameters

Bad quality SNPs were filtered by choosing parameters including Base Call Quality (<30), Genotype (1/1), and Mapping Quality (<50). The Base Call Qualities were calculated according to a Phred-scaled probability that the polymorphism exist at a certain site, and is expected to be low if the default threshold was not passed. All parameters concerning genotyping were excluded, although the 'Genotype' parameter showed for most SNPs a 1/0. If the given genotype was different (1/1) the SNP call was of bad quality. Further, the Mapping Quality parameter was included expressing the probability that an alignment to a given position in the reference genome is correct. As we did not have a consistent reference genome, the RMS Mapping Quality scores were not very high. The minimal mapping quality parameter was set to 50%, reflecting to the SNPs with highest quality scores. As extra control, we checked also the mapping quality zero parameters. They only passed the filter if no mapping quality zeroes were counted, which means that the particular SNP record was not covered by reads having a mapping quality of 0. After selection from the fixed parameters ~30% SNPs remained for analysis with variable parameters.

SNP selection - Variable parameters

In the first phase of SNP selection, the three parameters 'homopolymer run', 'read depth' and 'SNP ratio' were limited. The read depth limit was based on the estimated average coverage of 40x in the genome assembly. For the reduction of PCR-errors the length of homopolymers surrounding a SNP was set to a maximum of two nucleotides. To avoid misassembled SNPs and again PCR-errors, a coverage range of 30-60x was chosen. As this coverage parameter will not entirely filter out false positive SNPs, the SNP ratio parameter was also included. This SNP ratio is the division between the 'reference' nucleotide and the 'alternative' nucleotide. A SNP ratio of 1.0 was considered to be trustworthy, as this ratio shows a 50:50 appearance of respectively the reference and alternative SNP. Although this was a very stringent selection –as SNPs can also exist in e.g. a 20:80 ratio- the filtered list contained 84 SNPs that seemed to be consistent for genotyping in our *Asobara* populations.

In the second phase of SNP selection we have tried to expand the SNP list. However, we discovered also that our initial chosen parameters were not optimal. A graph of the total coverage against the number of bases learned that the median and the surrounding quartiles were a better estimate for our cut-off values. Therefore, it was chosen to set the read depth parameter to 30-110x in the new filter step.

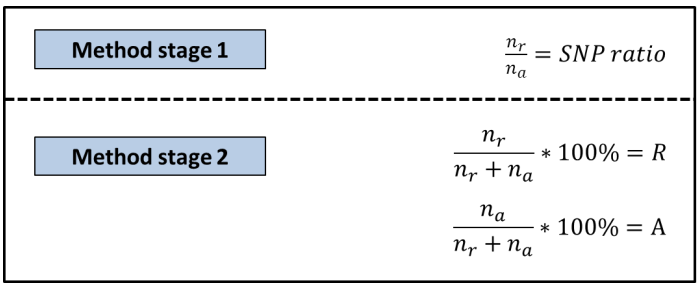

Figure 1 Calculations of the SNP ratio, in which n_x represents the number of reference resp. alternative nucleotides. As the 1st method did not meet equal under- and upper deviations, we have adapted the formula according to the 2nd method, in which R (reference) and A (alternative) are proportions of the total SNP coverage.

Further, the calculation of the SNP ratio was optimized. The ratio used during the first phase selection could not be used for other SNP proportions than 50:50. In SNP percentages of the reference and alternative bases (*R* and *A* respectively), we have corrected for the total nucleotides covering the SNP position (Figure 1). The optimal ratio compared to the gain of SNP counts is around 70:30 (SUPPLEMENTARY FILES Figure C-3). However, the main focus was pointed to SNPs that were positioned at the ends of the three largest scaffolds, in which this ratio was rather fluctuating. Without ratio filtration settings a list of 153 SNPs remained, of which 2 SNPs per scaffold were selected.

Blasting SNP primers

For molecular SNP detection, primers were designed flanking the 84 SNPs detected in the first phase of *in silico* SNP filtration. The primer design failed for 5 SNPs, as locations for proper annealing were lacking. The remaining SNP primers were checked by local nucleotide blasting against the reference genome (BLASTn data not shown). At least ~10% of the SNP primers were matching multiple times (>10) in its entire length. Another ~30% did anneal within 93% of its length. This remarkable high number of blast hits at some scaffolds indicates either a low quality assembly or presence of repetitive sequences.

SELECTION STEP	# SNPS (HKXTMS)	COMMENTS
Selected from VCF file	**84**	*69 scaffolds with 1 SNP* *6 scaffolds with 2 SNPs* *1 scaffold with 3 SNPs*
	- *37*	*Failed SNP primers BLASTn search*
Primer pairs covering SNP area	**47**	*Passed SNP primers BLASTn search*
No PCR product	- *10*	
PCR products (working primers)	**37**	
No sequence product	- *4*	
Homozygous SNPs	- *16*	Genotyped in F1 female
Heterozygous SNPs	**17**	
Homozygous (sequence error in F1)	- *3*	
Double nucleotide call at SNP position	- *4*	Genotyped in 50 F2 males
SNPs used for linkage grouping	**10**	

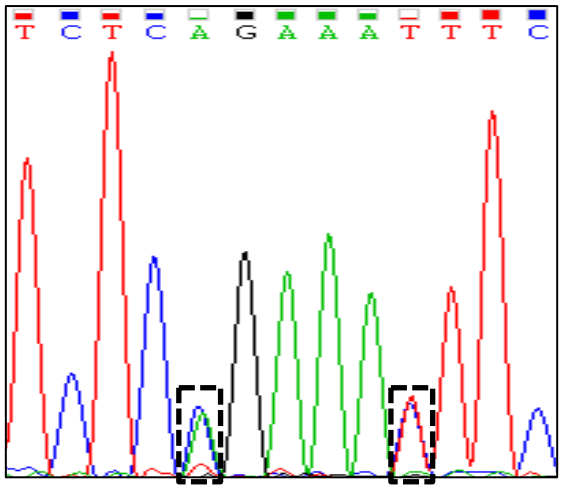

Figure 2 Table overview of SNPs dropped out during analysis. Four heterozygous SNPs (arrow) showed a double nucleotide in the haploid F2 individuals. The dashed boxes in the chromatogram mark the undefinable genotypes at SNP 11a (A/C) and 11b (C/T). These double peaks were found consistently in haploid DNA of the *A. tabida* inbred lines.

Molecular SNP analysis

47 SNP primer pairs were suitable for PCR, however, the product quality caused that genotyping was performed on 38 pairs only. The genotyping in the hybrid strains HKxTMS-3B showed that the F1 generation (virgin female) contained a total number of 17 heterozygous and 21 homozygous SNPs. All heterozygous SNPs were genotyped in the haploid F2 generation, resulting in a heterozygous:homozygous:indefinable number of respectively 10:3:4. Details of the genotyping analysis were reported in SUPPLEMENTARY FILES D. The chromatograms of the indefinable SNPs showed double peaks (Figure 2), which is contradictory to a haploid DNA strain. Both a nucleotide and protein blast against General-, Bacterial-, and *Wolbachia* databases did not confirm the origin of these of these odd SNP areas. The phenomenon was seen for SNPs 9, 11, 39, and 46 (SUPPLEMENTARY FILES Table D-2).

Maximizing the heterozygous genotyping of SNPs

Genotyping one HKxTMS line yielded only 10 heterozygous SNPs (Figure 3-1). To increase this number we have analysed the amount of heterozygous SNPs in three strains: TMS (2 inbred females), HKxTMS (3 hybrid females), and TMSxHK (1 hybrid female). It seemed that the mutual distribution of heterozygous SNPs was fluctuating between the different hybrid strains, however, no extraordinary gain was found in heterozygous SNP count. Genotyping showed that the TMS strain contained almost entirely homozygous SNPs (SUPPLEMENTARY FILES D2). Using different hybrid strains did not have a notable effect on the increase of the heterozygous SNP count. Because there was no extraordinary gain in SNP count in the other *Asobara* lines, it was decided to skip genotyping the F2 generations.

Construction of linkage groups

For the first estimation of linkage groups, the genotyping results of 10 SNPs were added in JoinMap®. Calculations in JoinMap® were based on a test of independence, which is translated into a LOD score. The principles of LOD-scaling are explained in SUPPLEMENTARY FILES E. The final grouping was done at several linkage thresholds of these LOD scores. In the first map the threshold was set at LOD=2.0. Although the constructed linkage map was not significant, it was notable that 5 out of 10 SNPs were grouped in one map (Figure 3). Even when the LOD-threshold was upgraded to the significant level of LOD=8.0, the linkage group includes the same 5 SNPs. The other 5 SNPs were ungrouped, suggesting that they were unlinked. This might indicate that the *A. tabida* genome contain at least one large chromosome compared to the other ~17 chromosomes. However, this might be a cautious estimation of groups and we have to keep in mind that the distance covered in this large linkage group is only 1.61‰ of the total genome length.

(1)

SCAFFOLD ID	SCAFFOLD SIZE (BP)	SNP ID	SNP (REF/ALT)	SNP POSITION
846	**18462**	**5a**	**A/G**	**17135**
		5b	**G/A**	**17186**
1041	16866	6	C/T	5684
1106	**16464**	**7**	**A/G**	**5906**
2652	11135	12a	T/C	5593
		12b	T/A	5621
3304	17283	14a	C/T	14693
		14b	C/G	14705
3472	**9784**	**17a**	**G/A**	**9666**
		17b	**G/T**	**9674**
11981	4854	26	G/A	4118
20580	2947	32	G/T	2454
30323	**1890**	**40**	**A/C**	**859**
36060	**1477**	**41**	**G/A**	**699**
3	140079	52a	C/T	14964
		52b	G/A	14907

(2)

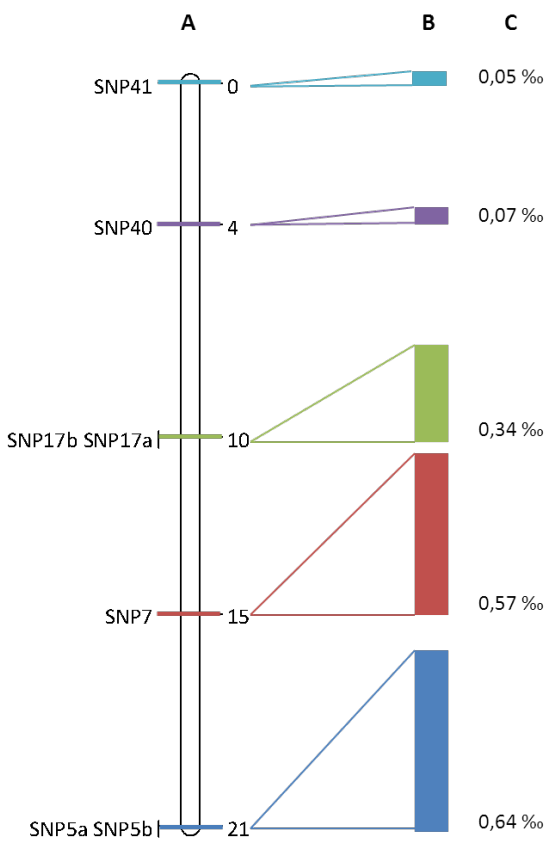

Figure 3 (1) Table overview of heterozygous SNPs used for linkage grouping. SNP 52 was originated from the second selection and not included in grouping analysis. SNPs marked in bold were placed in the same linkage group. Further details are shown in APPENDIX Table D-1 (2) Map of the estimated single linkage group (A) with LOD-score = 8.0. The scaffolds in this group had a cumulative length of ~0.05Mb, which was 1.61‰ of the total genome length (~300 Mb). The bars represent the proportion of this cumulative length per scaffold (B), whereas the permillage of the total genome length per scaffold was shown in (C).

Estimation of genetic distances

To estimate the mapping distances and the corresponding number of bases, 6 extra SNPs were included that are positioned at the ends of the three largest scaffolds. The registration of their recombination events may give information about their mutual distance on the chromosome. In this manner we can also check whether the scaffold was assembled correctly. Unfortunately we were not able to check the linkage, as 5 out of 6 SNPs were genotyped homozygous in the F1 females. The single heterozygous SNP (Figure 3-1) was not genotyped in the F2 generation, as more 'scaffold end' markers would be needed for estimating the SNP distances.

Discussion

In this study we have found many SNP markers in the genome of a highly inbred *Asobara* strain. As the first assembly of the sequenced reads resulted in more than 600000 scaffolds, we hoped we could collect an adequate number of SNPs to order several scaffolds into linkage groups.

Parameters in SNP filtration

The cumulative length of SNP containing scaffolds covered 70% of the total estimated genome length of ~300 Mb. Despite the relatively high coverage median of 63x, many SNPs were incorrectly called. In general, SNP quality assessments are performed using reference data. It includes parameters like -for example- transition/transversion (Ts/Tv) ratios and minor allele frequencies, and compares the called SNPs to the already reported SNPs in de SNP database (dbSNP) (White et al. 2007; Nielsen et al. 2011). Because the *A. tabida* genome was assembled for the first time, these parameters were not available. Parameter options that we could use in our SNP quality assessment were coverage (read depth), base call qualities, and mapping qualities. Initially, the first settings for read depth were based on the average coverage of 40x. This threshold was minimally chosen, as in a later stage the median –which was a better indicator of overall coverage- was calculated to be at ~60x. This setting had no direct consequences for SNPs with lower coverage. On the other hand, 25% of the SNPs with their coverage between 60-110 were excluded which means that the list of 84 SNPs was longer if we this depth boundary was collapsed from 60x to 110x.

The presence of homopolymers around the SNP location can interfere the SNP calling. In a different SNP calling strategy homopolymers larger than 4 were excluded (Ratan et al. 2010). In the first stage of SNP selection, the risk of homopolymer errors was reduced by exclusing SNP calls in homopolymer runs >2. It was suggested that the reliability of these 'homopolymer' SNPs will increase as the coverage threshold goes upwards (Komar 2009). However, the chance of assembly errors due to repetitive sequences and duplicated genes will rise if coverage is more than twice the average. This shows the importance of the coverage threshold setting to minimize both homopolymer ánd assembly errors. Nevertheless, we have tried to minimize these homopolymer errors in the second stage analysis, and excluded all SNPs located within homopolymers. Potential unbiased SNPs positioned in these areas were eliminated with this condition. As we were searching for high quality SNPs, the quality of homopolymer handling has to be assessed. This assessment might give insight into consequences of these homopolymer stretches on SNP quality. Due to time limitations, unfortunately we were not able to test the effect of homopolymer elimination on SNP heterzygosity in the inbred lines.

The mapping quality of the SNP list was lower (<60) compared to SNP calls derived with reference data (>90), which is a direct consequence of the completeness of read mapping. A simulation study that tested an assembly builder program (MAQ), showed that mapping qualities larger than ~50-60 in paired-end data are accurate enough for mapping based applications (Li et al. 2008). In our study, it was chosen to set the minimal mapping quality as a fixed parameter at 50%.

Although we have optimized the SNP ratio formula (Results, Figure 1), we have not constructed a new SNP list with ratio cut-offs. For this pilot study the 'old' ratio setting of strict 50:50 was sufficient to generate a global idea of SNP distribution. However, SNPs are present in different ratios and therefore we may have missed many SNPs useful for genotyping and mapping. According to the cut-off tests for mapping quality and read depth ratio (SUPPLEMENTARY FILES Figure C2 and C3), there is much fluctuation in the output of SNPs. For future research, it is recommended to further test and optimize these parameter settings and their influence on true SNP calls.

Judging duplicated fragments in de novo genome assembly

When the primers were blasted against the reference genome, a remarkable high number of 'mismatches' or double alignments were present. Multiple annealing sites will make unambiguous amplification more difficult and therefore ~40 out of 84 SNPs were excluded from further analyses. The presence of these double aligned primers might criticize the assembly quality. Either the genome consists of numerous repetitive parts or the assembler separated highly similar sequence strands incorrectly. The ABySS assembler used a two-stage algorithm. The first stage concerned the filtration of read errors and contig building, whereas the second performs contig extension by including paired-end information (Simpson et al. 2009). Especially in this last stage either collapsing or merging of highly similar sequence copies is not perfectly balanced, which is a known problem in the current assembly algorithms. (Kelley and Salzberg 2010). We have tried to reduce mis-assembled reads by applying the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009). During this second mapping procedure it was considered whether reads mate pairs covering highly similar sequences are more consistent if duplicated reads are either merged or collapsed. Both the results of BLAST and of the 'heterozygous SNPs' in the haploids are pointing to these problems. The output of ABySS was used as a 'consensus' sequence for secondary mapping. Therefore, it is most likely that assembly errors occurred during the ABySS alignment. Up to now, unfortunately no clear solution was found for this problem.

Sample size improves quality of LG estimation

The recombination fraction is a measure for the number of recombinants and is 0,5 at maximum. On a genetic map this corresponds with a distance of 50 cM, and is the maximum distance we can measure between two markers (Strachan and Read 1999; Klug and Cummings 2002). For our study we started with genotyping 50 individuals per SNP. Initially this amount is the absolute minimum for linkage grouping. If we consider a linkage significance level of p=0.05 (Figure 4) the maximal genetic distance that we can measure between two loci is 38 cM. This distance increases when more individuals are genotyped, which has the advantage that the chance of finding informative SNPs will increase. In our case, doubling the amount of genotyped individuals to n=100 would gain the significant genetic distance with 'just' 4 cM. For this pilot study the choice of n=50 was appropriate for setting up a linkage mapping strategy. However, for constructing a proper linkage map with more accurate genetic distances it might be better to genotype at least 200 individuals per SNP.

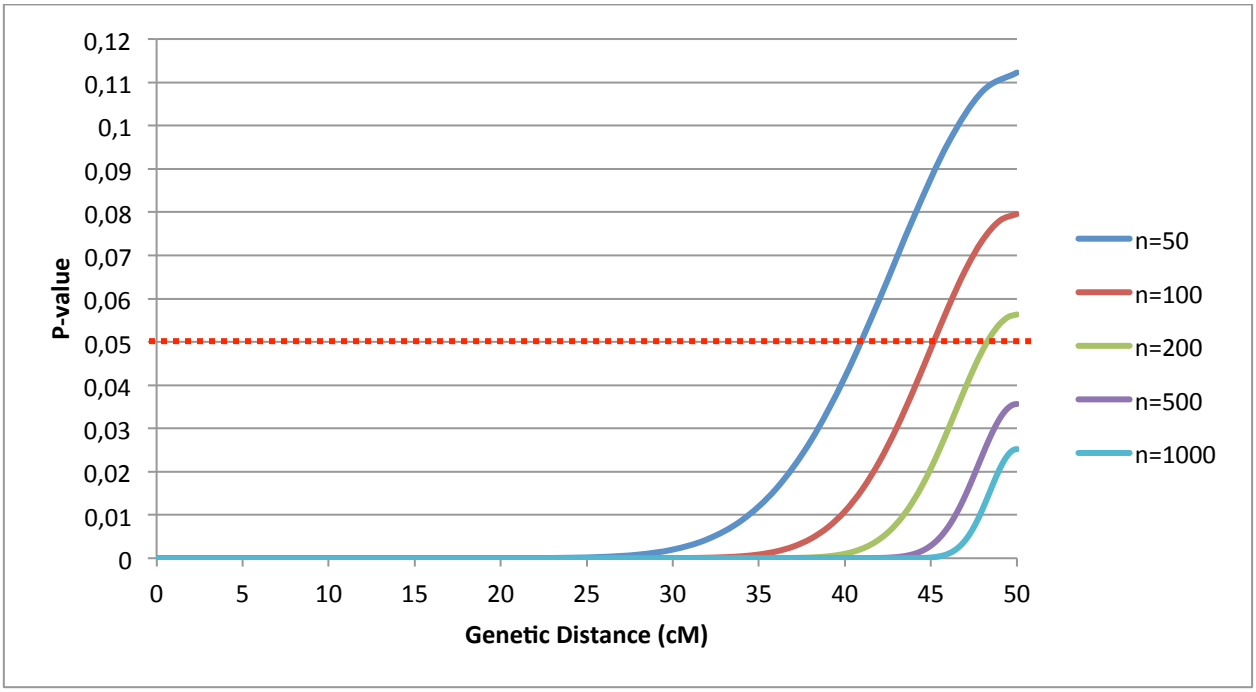

Figure 4 Binomial distributions for significant linkage, accounted per sample size *n*. The red dotted line marks the significant p=0.05 threshold.

Linkage grouping in inbred lines

It was expected that a minimum of 80 SNPs was needed to construct appropriate linkage groups. Due to time limitations of this thesis, we were able to find 10 SNPs and constructed one large linkage group of 5 SNPs. Based on the information of haplotypes, the other 5 SNPs seemed to be unlinked. A general accepted minimal LOD-score for linkage is 3.0. The linkage group we have found had a LOD-score of 8.0, which corresponds to strong linkage. However, when multiple markers indicate linkage,

the LOD threshold should be corrected for the involved number of markers (*n*), which corresponds to 3+log(*n*) (Strachan and Read 1999). In our linkage model, 5 SNP markers were involved in one group, thus the threshold should change into LOD ≥ 3.7. Even with this very stringent threshold the group showed strong evidence for linkage. Nevertheless, it is possible that these SNPs are positioned at dependent segregating chromosomes, as we analysed DNA of inbred lines only. Genes (or fragments) are not inherited at random in inbred populations. The number of analysed SNPs is too low to distinguish if SNPs are present at similar or different chromosomes. The general accepted number of SNP markers needed for linkage grouping is calculated at 5 SNPs per chromosome (Kearsey and Pooni 1996). Assuming 17 chromosomes for *Asobara* species, at least 85 heterozygous SNPs are needed for adequate linkage grouping. A simple calculation points out that we need at least 714 SNPs from *in silico* SNP filtration to get a minimum of 85 SNPs per analysed hybrid female.

Conclusion

Our method succeeded in finding SNPs in inbred lines, although the progress of genotyping was less efficient. Linkage group analysis of a few genotyped SNPs resulted in one large group compared to separate (unrelated) SNPs. This may indicate high variation in chromosome sizes in which at least one large chromosome is present. However, for proper estimation of the chromosome number we need to include more SNPs in our analysis. If the number of filtered SNPs will increase, e.g. by optimizing the mapping quality and SNP ratio values, more SNPs could be genotyped as heterozygous and useful for mapping. Nevertheless, it is rather time consuming to genotype this large amount of SNPs. Therefore, it is recommended to analyse SNPs in the DNA of outbred strains in future. In this manner the chance for finding heterozygous SNPs will considerably increase. Further, karyotyping of the *Asobara* genome might help in establishing the chromosome number and sizes.

Acknowledgements

I would like to thank Leo Beukeboom for the opportunity to accomplish my thesis at the group of Evolutionary Genetics in Groningen. Special thanks go to my supervisors, Eveline Verhulst, Pieter Neerincx, and Louis van de Zande, for the interesting work discussions and their help and support during practical and theoretical issues. Thanks to everyone of the group for your help, support, work discussions and social activities. I have had a great time!

References

Aljanabi, S. M. and I. Martinez (1997). "Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques." Nucleic acids research 25(22): 4692-4693.

Beukeboom, L. W., J. Ellers and J. J. Van Alphen (2001). "Absence of single-locus complementary sex determination in the braconid wasps Asobara tabida and Alysia manducator." Heredity 84(1): 29-36.

Beukeboom, L. W., A. Kamping and L. van de Zande (2007). "Sex determination in the haplodiploid wasp Nasonia vitripennis (Hymenoptera: Chalcidoidea): a critical consideration of models and evidence." Seminars in cell & developmental biology 18(3): 371-378.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth and S. T. Sherry (2011). "The variant call format and VCFtools." Bioinformatics 27(15): 2156-2158.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas and M. Hanna (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nature genetics 43(5): 491-498.

Geuverink, E. (2011). "Genetics of sex determination in haplodiploid wasps." Introductory Essays in Funtional Ecology (CEES) 84: 41.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series.

Kearsey, M. J. and H. S. Pooni (1996). The genetical analysis of quantitative traits. London [etc.], Chapman & Hall.

Kelley, D. R. and S. L. Salzberg (2010). "Method Detection and correction of false segmental duplications caused by genome mis-assembly."

Klug, W. S. and M. R. Cummings (2002). Essentials of Genetics, Prentice Hall.

Komar, A. A. (2009). Single nucleotide polymorphisms : methods and protocols. [New York], Humana Press.

Kraaijeveld, A. R., I. C. T. Adriaanse and d. B. B. van (1999). "Parasitoid size as a function of host sex: potential for different sex allocation strategies." Entomologia Experimentalis et Applicata 92(3): 289-294.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics 25(14): 1754-1760.

Li, H., J. Ruan and R. Durbin (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res 18(11): 1851-1858.

Ma, W.-J., B. Kuijper, J. G. de Boer, L. van de Zande, L. W. Beukeboom, B. Wertheim and B. A. Pannebakker (2013). "Absence of Complementary Sex Determination in the Parasitoid Wasp Genus <italic>Asobara</italic> (Hymenoptera: Braconidae)." PLoS ONE 8(4): e60459.

Marshall, O. J. (2004). "PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR." Bioinformatics 20(15): 2471-2472.

Milne, I., M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright and D. Marshall (2010). "Tablet--next generation sequence assembly visualization." Bioinformatics 26(3): 401-402.

Morin, P. A., G. Luikart, R. K. Wayne and S. N. P. w. g. the (2004). "SNPs in ecology, evolution and conservation." Trends in ecology & evolution (Personal edition) 19(4): 208-216.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song (2011). "Genotype and SNP calling from next-generation sequencing data." Nat Rev Genet 12(6): 443-451.

Ratan, A., Y. Zhang, V. M. Hayes, S. C. Schuster and W. Miller (2010). "Calling SNPs without a reference sequence." BMC bioinformatics 11(1): 130.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones and I. Birol (2009). "ABySS: a parallel assembler for short read sequence data." Genome Res 19(6): 1117-1123.

Strachan, T. and A. P. Read (1999). "Genetic mapping of Mendelian characters."

Van Oijen, J. and R. Voorrips (2001). "Joinmap Version 3.0, software for the calculation of genetic linkage maps." Plant Research International, Wageningen, The Netherlands.

Van Wilgenburg, E., G. Driessen and L. W. Beukeboom (2006). "Single locus complementary sex determination in Hymenoptera: an "unintelligent" design." Front. Zool 3(1).

White, T. L., W. T. Adams and D. B. Neale (2007). Genetic markers - morphological, biochemical and molecular markers. Forest Genetics. CABI: 53-76.