Chair of the examination committee:

Prof. Dr. Godelieve Gheysen

Prof. Dr. ir. Frank Devlieghere (replacing chair)

Members of the reading committee:

Prof. Dr. ir. Tim De Meyer

Prof. Dr. Filip Volckaert

Prof. Dr. ir. Guy Smagghe

Dr. Ole Madsen

Other members of the examination committee:

Prof. Dr. Patrick Sorgeloos

Stephane Rombauts

Promoters (Ghent University):

Prof. Dr. ir. Peter Bossier

Dr. ir. Marnik Vuylsteke

Dean of the Faculty of Bioscience engineering of Ghent University

Prof. dr. ir. Guido Van Huylenbroeck

Ghent University Rector

Prof. dr. Anne De Paepe

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

Faculty of Bioscience Engineering (FBE)

Department of animal production

Faculty of Sciences

Department of Plant Biotechnology and Bioinformatics

VIB Department of Plant Systems Biology

Genomic tools and sex determination in the extremophile brine shrimp *Artemia franciscana*

Stephanie De Vos

Promoters: Prof. Dr. ir. Peter Bossier and Dr. ir. Marnik Vuylsteke

Thesis submitted in fulfillment of the requirements for the degree of

Doctor (PhD) in Applied Biological Sciences

Academic year: 2013-2014

LABORATORY OF AQUACULTURE
ARTEMIA REFERENCE CENTER

Plant Systems Biology

Genomische tools en seks determinatie in het extremofiele pekelkreeftje *Artemia franciscana*

Copyright © Stephanie De Vos, Faculty of Bioscience engineering, Ghent University, 2014. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without permission in writing from the copyright holder.

Cite as:
De Vos, S (2014) Genomic tools and sex determination in the extremophile brine shrimp *Artemia franciscana*. PhD thesis, pp 157.

Summary

The aim of this study was the construction of a genomic *Artemia* toolkit. Sex-specific AFLP-based genetic maps were constructed based on 433 AFLP markers segregating in a 112 full-sib family, revealing 21 male and 22 female linkage groups ($2n = 42$) covering 1,041 and 1,313 cM, respectively. Fifteen putatively homologous linkage groups, including the sex linkage groups, were identified between the female and male linkage maps. Eight sex-linked markers, heterozygous in female animals, mapped to a single locus on a female linkage group, supporting the hypothesis of a WZ/ZZ genetic sex-determining system and showing primary sex determination is likely directed by a single gene. The produced *Artemia* linkage maps provide the basis for further fine mapping and exploration of the sex-determining region and are marker resources for mapping genomic loci underlying phenotypic differences among *Artemia franciscana* strains.

To fine-map the sex locus, bulked segregant analysis was performed: next-generation sequencing reads were generated from one male and one female pool of full-sib *A. franciscana*. The sequences were assembled *de novo* to an *A. franciscana* draft genome, and male and female reads were mapped onto the draft genome. Scaffolds containing genes with SNPs heterozygous in females and homozygous in males were selected and analyzed for gene content. Candidate primary sex-determining genes were identified, including *Cytochrome P450*, *F0F1 ATP synthase subunit beta*, a gene containing a CRAL-TRIO domain and a gene containing an ankyrin repeat, two *Fibronectin* genes and finally *SEC14* and *Zinc finger C2CH-type*. Several genes, highly homologous with insect sex-determining genes (*doublesex*, *sex-determining region Y*, *sex-lethal*, *feminization 1*, *transformer 1*, *transformer 2*, *fruitless*, *runt*, *deadpan*, *daughterless*, *extra macrochaetae*, *groucho* and *sans fille*) or with crustacean sex-related genes (*extra macrochaetae*, *fushi tarazu*, *forkhead box*, *WNT*, *argonaute*, *testis-specific*, *VASA*, *SOX*, *star*, *ECM*, *tudor*, *GATA*, *prohibitin* and *Cytochrome P450*) were found in the *A. franciscana* draft genome. Of these genes, only *Cytochrome P450*, which, through transcriptomic studies, is already known as a candidate sex-determining gene for *Macrobrachium nipponense*, was selected by BSA, indicating that it may play a role in primary sex determination in *Artemia.*

The 1,310-Mbp *Artemia* draft genome sequence (N50 = 14,784 bp; GC-content = 35%; 176,667 scaffolds) was annotated, predicting 188,101 genes with an average length of 692 bp (Chapter 4).

Ninety percent of the *Artemia* ESTs available on NCBI, as well as 92.2% of the RNAseq reads from the transcriptome of cysts at different metabolic stages (anoxia, diapause, quiescence and hydration) and from larvi kept at different salinities (30 g/l and 200 g/l) were present in the *Artemia* genome, indicating that the functional part of the genome under the RNAseq sampling conditions is virtually fully represented in the assembly.

Several steps were taken in this study to introduce *Artemia* as a new genomic model for crustaceans. Further testing of the candidate primary sex-determining genes should narrow down the selection towards one primary sex-determining gene for *Artemia*. Although the functional part of the *Artemia* genome under the RNAseq sampling conditions is virtually fully represented in the assembly, thus making it useful for qualitative research, genome finishing strategies will still be necessary to complete the genome project. The further development of genomic resources for *Artemia* will add a completely new dimension to *Artemia* research and its use as live food in aquaculture.

Samenvatting

De doelstelling van deze studie was het creëren van een genomische toolkit voor *Artemia*.

Seks-specifieke AFLP-gebaseerde genetische kaarten van respectievelijk 1041 en 1313 cM lang werden geconstrueerd op basis van 433 AFLP merkers die segregeren in een 112 full-sib familie en 21 mannelijke en 22 vrouwelijke koppelingsgroepen (2n = 42) werden geïdentificeerd. Vijftien vermeende homologe koppelingsgroepen, met inbegrip van de seks koppelingsgroepen, werden geïdentificeerd tussen de vrouwelijke en mannelijke koppelingskaart. Acht geslachtsgekoppelde merkers, heterozygoot in vrouwelijke dieren, karteerden op één enkel locus op een vrouwelijke koppelingsgroep. Dit ondersteunt de hypothese van een WZ/ZZ genetisch seks-determinerend systeem en toont aan dat primaire geslachtsbepaling waarschijnlijk wordt bepaald door één gen. De geproduceerde genetische kaarten voor *Artemia* vormen de basis voor het verder fijnkarteren en onderzoeken van het seks-determinerende gebied en zijn een bron van merkers voor het in kaart brengen van genomische loci die aan de basis liggen van fenotypische verschillen tussen *Artemia* soorten.

Om het sex locus te fijnkarteren werd Bulked Segregant Analysis uitgevoerd: Next Generation Sequencing sequenties van een mannelijke en een vrouwelijke pool van full-sib *A. franciscana* werden gegenereerd, de sequenties werden *de novo* geassembleerd tot een *A. franciscana* genoom en mannelijke en vrouwelijke sequenties werden gemapt op het genoom. Genoomfragmenten die genen bevatten met SNPs, heterozygoot in wijfjes en homozygoot in mannetjes, werden geselecteerd en verder geanalyseerd voor gen-inhoud. Kandidaat primair seks-determinerende genen werden geïdentificeerd, waaronder *Cytochrome P450*, *F0F1 ATP synthase subunit beta*, een eiwit dat een CRAL-TRIO domein bevat, een eiwit dat ankyrin repeats bevat, twee *Fibronectin* genen, *SEC14* en *Zinc finger C2CH-type*. Verschillende genen, sterk homoloog met seks-determinerende genen van insecten (*doublesex*, *sex-determining region Y*, *sex-lethal*, *feminization 1*, *transformer 1*, *transformer 2*, *fruitless*, *runt*, *deadpan*, *daughterless*, *extra macrochaetae*, *groucho* en *sans fille*) of met seks-gerelateerde genen van schaaldieren (*extra macrochaetae*, *fushi tarazu*, *forkhead box*, *WNT*, *argonaute*, *testis-specific*, *VASA*, *SOX*, *star*, *ECM*,

tudor, *GATA*, *prohibitin* en *Cytochrome P450*) werden gevonden in het *A. franciscana* genoom. Van deze genen werd enkel *Cytochrome P450* geselecteerd door BSA, een gen dat via transcriptoomstudies een bekend kandidaat seks-determinerend gen van schaaldier *Macrobrachium nipponense* is, wat aangeeft dat het een rol kan spelen in de primaire geslachtsbepaling in *Artemia*.

Het 1310 Mbp grote *Artemia* genoom (N50 = 14.784 bp; GC-gehalte = 35%; 176.667 fragmenten en 1310 Mbp) werd geannoteerd en 188.101 genen met een gemiddelde lengte van 692 bp werden voorspeld. Negentig procent van de Artemia ESTs op NCBI en 92.2% van de RNAseq sequenties van het transcriptoom van cysten in verschillende metabolische toestanden (anoxie, diapause, quiescentie en gehydrateerd) en van larven gehouden bij verschillende zoutgehaltes (30 g/l en 200 g/l) waren aanwezig in het *Artemia* genoom, wat aantoont dat het functionele deel van het genoom in de omstandigheden van RNAseq staalname vrijwel volledig weergegeven is in het genoom.

Door dit onderzoek staat *Artemia* al een stap dichter bij het worden van een nieuw genomisch model voor schaaldieren. Verder onderzoek van de kandidaat primair seks-determinerende genen zullen de selectie beperken tot één primair seks-determinerend gen voor *Artemia*. Hoewel het functionele deel van het *Artemia* genoom onder de condities van staalname voor RNAseq vrijwel compleet is vertegenwoordigd in het genoom, waardoor het nuttig is voor kwalitatief onderzoek, zijn strategieën om het genoom af te werken nog steeds noodzakelijk om het genoom project te voltooien. De verdere ontwikkeling van genomische tools voor *Artemia* zal een geheel nieuwe dimensie toe te voegen aan *Artemia* onderzoek en het gebruik ervan als levend voer in de aquacultuur.

Table of contents

ACKNOWLEDGEMENTS XIII

LIST OF FIGURES AND TABLES XV

LIST OF ABBREVIATIONS XIX

CHAPTER 1

INTRODUCTION

1.1. *Artemia* 1

1.1.1. Taxonomy 1

1.1.2. Habitat 4

1.1.3. Life cycle and extremophilic features 5

1.1.4. Applications in the aquaculture industry and in research 6

1.2. Genomics 8

1.2.1. Genomics and gene discovery: main techniques used 8

1.2.1.1. Genome mapping 8

1.2.1.2. Bulked segregant analysis 10

1.2.1.3. From organism to genome 12

-a) Sequencing 12

-b) Assembly 16

-c) Quality characteristics of a *de novo* genome assembly 18

-d) Genome Annotation 19

-e) NGS genomes: facts to keep in mind 21

1.2.2. Arthropod genome projects 21

1.2.3. *Artemia* genomics 22

1.3. Sex determination and sex-determining genes in arthropods........................24
1.4. Research aims...30

CHAPTER 2

A FIRST AFLP-BASED GENETIC LINKAGE MAP FOR BRINE SHRIMP *ARTEMIA FRANCISCANA* AND ITS APPLICATION IN MAPPING THE SEX LOCUS

2.1. Abstract 33
2.2. Introduction..34
2.3. Materials and methods 36
  2.3.1. Mapping population 36
  2.3.2. DNA extraction 36
  2.3.3. Segregation analysis and linkage mapping....................................36
  2.3.4. *Artemia* genome size estimation by flow cytometry 38
2.4. Results 40
  2.4.1. Segregation analysis and linkage mapping....................................40
  2.4.2. Linkage mapping of the sex locus..45
  2.4.3. *Artemia* genome size estimation by flow cytometry 46
2.5. Discussion 48

CHAPTER 3

FURTHER CHARACTERIZATION OF THE SEX-DETERMINING REGION IN *ARTEMIA*

3.1. Abstract 51
3.2. Introduction..52
  3.2.1. Importance of sex-determining genes 52
  3.2.2. Sex determination mechanisms in arthropods 52
    3.2.2.1. Insects..52
    3.2.2.2. Crustaceans 52
3.3. Materials and methods 55

3.3.1. Rationale for potential *Artemia* sex determination systems 55

-1) The primary sex-determining gene lies on the Z chromosome 55

-2) The primary sex-determining gene lies on the W chromosome.. 56

-3) The primary sex-determining gene lies on both the W and Z chromosomes. 57

3.3.2. AFLP analysis, sequencing and BLAST of sex-linked markers........................ 58

3.3.3. BSA by whole-genome sequencing 58

3.3.4. Sex-specific genes with non-synonymous SNPs... 62

3.3.5 Arthropod sex-determining gene homologs, not selected by BSA.. 63

3.4. Results 64

3.4.1. AFLP analysis, sequencing and BLAST of sex-linked markers........................ 64

3.4.2. Candidate sex-determining genes selected by BSA 68

3.4.3. Sex-specific genes with non-synonymous SNPs... 75

3.4.4. Arthropod sex-determining gene homologs, not selected by BSA.. 79

3.5. Discussion 81

3.5.1. AFLP analysis, sequencing and BLAST of sex-linked markers........................ 81

3.5.2. The sex-determining system of *Artemia* 81

3.5.3. Bulked segregant analysis by whole-genome sequencing 81

3.5.4. Arthropod sex-determining gene homologs, not selected by BSA.. 85

3.5.5. *Artemia* candidate sex-determining genes 86

CHAPTER 4

THE *ARTEMIA* GENOME

4.1. Abstract 87

4.2. Introduction... 88

4.3. Materials & Methods 89
  4.3.1. Samples for sequencing 89
    4.3.1.1. PE and MP DNA 89
    4.3.1.2. RNA 89
  4.3.2. Sequencing 93
    4.3.2.1. PE DNA 93
    4.3.2.2. Cre-lox MP DNA 93
    4.3.2.3.PE RNA 93
  4.3.3. Genome assembly 94
  4.3.4. Genome annotation 96
  4.3.5. Homology of *L. vannamei* protein in the *Artemia* genome 98
4.4. Results 99
  4.4.1. Genome assembly 100
  4.4.2. Genome annotation 104
  4.4.3. Homology of *L. vannamei* protein in the *Artemia* genome 107
4.5. Discussion 108
  4.5.1. Genome assembly 108
    4.5.1.1. Fragmentation 108
    4.5.1.3. Repeats 108
    4.5.1.3. Completeness of functional sequence 110
  4.5.2. Genome annotation 111
  4.6. Future research 112

CHAPTER 5

DISCUSSION AND PERSPECTIVES

5.1. A genomic toolkit for *Artemia* 113
5.2. General discussion, future research and applications 114
  5.2.1. Linkage maps 114

5.2.2. Putative primary sex-determining genes 115

5.2.2.1. Detected by BSA 115

5.2.2.2. Arthropod sex-determining gene homologs, not detected by BSA 117

5.2.3. The genome 117

5.2.3.1. Fragmentation 117

5.2.3.2. Differential expression 120

5.2.3.3. Comparative genomics, pathway mapping and QTL analysis 121

5.3. Conclusion 122

REFERENCES.. 123

CURRICULUM VITAE 151

Acknowledgements

It is with great pleasure that I present this work to you. Of course, I have not made this accomplishment on my own. The funding for this work has been provided by the BOF grant supporting my PhD, the Laboratory of Aquaculture & *Artemia* Reference Center and VIB Techwatch. As most PhD students, I was supported by networks of professors, scientists, technicians, friends and family (categories overlap in several cases). That is why I would like to put them in the spotlight and thank all the people who helped me to climb the steps of the ladder towards this PhD.

Without the support of my promoters Prof. Dr. ir. Peter Bossier from the Laboratory of Aquaculture & *Artemia* Reference Center and Dr. ir. Marnik Vuylsteke from the Quantitative Genomics group in the VIB-PSB, this thesis would not have been a reality. Very closely involved in each step of the making of this thesis, Marnik Vuylsteke introduced me to the world of genomics, provided expert advice on genetics, relentlessly revised me through my first first-author article (represented by Chapter 2 of this work) until it was just right and helped me with his expertise on statistical analysis. Peter Bossier was a source of inspiration to take the research in new directions and for thoughts on genetics and *Artemia*. It was very useful to be challenged on many occasions by his out-of-the-box thinking. I think there is no more adventurous and rewarding way of doing science. Gilbert Vanstappen enlightened me with his expert view on *Artemia* populations, taxonomy and biology.

I also want to thank Ilse Vercauteren, who consistently gave me useful advice, mostly work-related, but also about personal matters. She taught me the AFLP technique and gave me the knowledge, gestures and tricks that make all the difference in achieving a successful experiment. She performed the RNA extractions used in this work as well. Anita De Haese carried out part of the large-scale RNA-seq experiments and Christ Mahieu advised and helped me with *Artemia* hatching and rearing. Gert Van Isterdael taught me how to use the MoFlo, a flow cytometer which looks like nothing less than a helicopter cockpit!

For the bioinformatics part of this thesis, including working in a UNIX environment, sequencing, assembling, mapping, annotating sequences and countless other topics, Stephane Rombauts was of invaluable advice and support. Stephane introduced me to a recent field of expertise that I will, no doubt, benefit from in the future as well. Frederik Delaere introduced me to the do`s and don`ts of working with a computer cluster. I climbed the rest of the bioinformatics learning curve with Ives Vanstechelman, my lab neighbor, who had a similar topic to mine, while working with another organism. We shared our insights and were a good team for work, but also for eating waffles, planting papyruses and enjoying lunch breaks with Valerie, Annelies and Mathias, Tom and Anna from the Programmed Cell Death group (with or without Frisbee).

I met VIB and ARC members at birthdays, PhD defenses, team buildings, congresses or one of the many VIB or ARC events. It was a great experience to work in such an international environment. I enjoyed all the conversations, laughs, dinners, concerts, weddings and trips we had together. To name a few: Chiarina Darrah, Eleonora Billi, Ives Vanstechelman, Tom Baelemans, Miguel Cardoso, Stephane Rombauts, Naira Naouar, Lorin Spruyt, the Programmed Cell Death group and many more. I owe a lot of fun but also, motivation, support and chocolate-driven thesis writing sessions to my great, creative and crazy friend Dorien Colman since we traveled to Cuba together for our master thesis.

A special thanks goes out to my family. My family has always been my first and closest support system. They helped me perform during these thesis years in spite of a series of personal hardships. I want to thank Pascaline De Vos, my sister, for providing me with sisterly advice, distractions and dog walks with Canelle during the thesis writing period. I would like to thank Claudine Bailleul, my mother, in particular. Relying on one income, she offered both her children a university education, by relentlessly juggling her rising career and raising us as a present mom at the same time. She deserves to feel proud about such an achievement. During this thesis, she has consistently motivated me at difficult times, invited me to hip activities in Brussels to relax, listened when I was going on and on about some shrimp tank/experiment/software/(insert other scientific topics here) and given me sound advice when I was in doubt. For that, I am grateful beyond words.

Stephanie

List of figures and tables

Figure 1-1. Pancrustacean phylogeny 2

Table 1-1. Taxonomy of the brine shrimp *Artemia* .... **Error! Bookmark not defined.**3

Figure 1-2. Phylogenetic relationships among *Artemia* species 4

Figure 1-3. The life cycle of sexually reproducing *Artemia* 5

Figure 1-4. Outline of the AFLP procedure... 9

Figure 1-5. Illumina sequencing 14

Figure 1-6. Basic steps in sequence assembly, scaffolding and genome finishing....... 16

Figure 1-7. Three basic approaches to genome annotation 20

Figure 1-8. Crustacean species selected for the i5k project.. 21

Figure 1-9. Circular physical map of the *Artemia tibetiana* mitochondrial genome.... 23

Figure 1-10. Simplified view of sex-determining pathways in Arthropoda................. 25

Figure 1-11. The X:A ratio determines sex in *Drosophila* 26

Figure 1-12. Sex determination cascade in *Drosophila* 27

Figure 1-13. Phylogeny of DM domain-containing genes**Error! Bookmark not defined.**28

Figure 1-14. Four *Artemia franciscana* karyotypes 29

Table 2-1. List of the 65 primer combinations used for AFLP analysis........................ 37

Figure 2-1. *A. franciscana* autosomal female linkage groups 41

Figure 2-2. *A. franciscana* autosomal male linkage groups.. 42

Table 2-2. Statistics for the female and male linkage maps.. 43

Figure 2-3. *Artemia franciscana* homologous autosomal female and male linkage groups 44

Figure 2-4. *Artemia franciscana* sex linkage groups.. 45

Table 2-3. Segregation patterns of sex-linked AFLP markers....................................... 45

Figure 2-5. Fluorescence histograms for *Artemia franciscana* and chicken DNA content estimation 47

Figure 3-1. A WZ/ZZ sex-determining system showing the the sex-determining gene present on the Z chromosome 55

Figure 3-2. A WZ/ZZ sex-determining system showing the sex-determining gene present on the W chromosome 56

Figure 3-3. A WZ/ZZ sex-determining system showing the sex-determining gene present on both the W and Z chromosomes as allelic variants 57

Figure 3-4. Distribution of the probability P of having k times an allele A, assuming a

SNP covered N times with alleles A and T 61
Table 3-1. AGP file for LG_Female_1, the new scaffold created by anchoring of SM-containing scaffolds according to the *Artemia* genetic linkage map 64
Table 3-2. BLAST results of six sex-linked AFLP markers on the *Artemia* genome (Scaffold_13).. 66
Figure 3-5. Scaffold_13, reference allele frequency of selected SNPs in female and male reads with respective spline plots and sex-linked AFLP markers 66
Figure 3-6. Selected SNPs and detected indels present in Scaffold_13 at the location of sex-linked AFLP marker E112M223M273.0F 67
Figure 3-7. Base coverage of scaffold_13 by male and female PE reads 68
Table 3-3. Genes with known function found on fragments selected by BSA, unrelated to sex determination or sex differentiation mechanisms 69
Table 3-4. Genes with known function found on fragments selected by BSA, rejected based on less clear SNP allele frequency differences in their respective spline plots. 70
Table 3-5. Genes with functional description linked to sex determination or differentiation 71
Figure 3-8. Scaffold_15143, reference allele frequency of selected SNPs in female and male reads with respective spline plots and a gene containing the CRAL-TRIO domain (artfr15143g00010) and *F0F1 ATP synthase subunit beta* (artfr15143g00020). 72
Figure 3-9. Scaffold_15484, reference allele frequency of selected SNPs in female and male reads with respective spline plots and the genes *Fibronectin, type III* (artfr15484g00020 & artfr15484g00030). 72
Figure 3-10. Scaffold_2087, reference allele frequency of selected SNPs in female and male reads with respective spline plots and a gene containing Ankyrin repeat (artfr2087g00020)... 73
Figure 3-11. Scaffold_21607, reference allele frequency of selected SNPs in female and male reads with respective spline plots and the gene *Zinc finger, C2CH-type* (artfr21607g00030). 73
Figure 3-12. Scaffold_23412, reference allele frequency of selected SNPs in female and male reads with respective spline plots and the gene *Cytochrome P450* (artfr23412g00030). 74
Figure 3-13. Scaffold_4146, reference allele frequency of selected SNPs in female

and male reads with respective spline plots and the gene *SEC-14-like protein 1* (artfr4146g00010). 74
Figure 3-14. Alignment of *A. franciscana* reference protein artfr15143g00010 (from a gene containing the CRAL-TRIO domain) and its variant protein 75
Figure 3-15. Alignment of *A. franciscana* reference protein artfr15143g00020 (from *F0F1 ATP synthase subunit beta*) and its variant protein. 75
Figure 3-16. Alignment of *A. franciscana* reference protein artfr15484g00020 (from *Fibronectin, type III*) and its variant protein 75
Figure 3-17. Alignment of *A. franciscana* reference protein artfr15484g00030 (from *Fibronectin, type III*) and its variant protein 75
Figure 3-18. Alignment of *A. franciscana* reference protein artfr2087g00020 (from a gene containing Ankyrin repeat) and its variant protein 76
Figure 3-19. Alignment of *A. franciscana* reference protein artfr23412g00030 (from *Cytochrome P450*) and its variant protein 76
Table 3-6. Synonymous and non-synonymous selected SNPs in the eight genes showing only detected SNPs, heterozygous in females and homozygous in males. 77
Figure 3-20. Changes in secondary protein structure due to non-synonymous SNPs in *CRAL-TRIO domain* (artfr15143g00010) 78
Figure 3-21. Changes in secondary protein structure due to non-synonymous SNPs in *F0F1 ATP synthase subunit beta* (artfr15143g00020) 78
Figure 3-22. Changes in secondary protein structure due to non-synonymous SNPs in a gene containing Ankyrin repeat (artfr2087g00020) 78
Figure 3-23. Changes in secondary protein structure due to non-synonymous SNPs in *Cytochrome P450* (artfr23412g00030) 79
Table 3-7. Genes homologous to known insect sex-determining genes found in the *Artemia* genome 79
Figure 3-24. Sex determination in *Drosophila* from the chromosomal signaling cascade until *sxl* and from *sxl* until *dsx* 80
Figure 4-1. Sampling of *A. franciscana* for RNA-seq 92
Table 4-1. *Artemia* sampling for RNAseq. 92
Figure 4-2. Assembly of the *Artemia* draft genome sequence 96
Figure 4-3 Annotation pipeline for the *Artemia* genome 98
Table 4-2. Libraries used for sequencing of DNA and RNA. 99

Table 4-3. Overview of key assembly values 100
Figure 4-4. Evolution of the number of scaffolds in the subsequent *Artemia* assembly stages. 101
Figure 4-5. Evolution of the N50 value in the subsequent *Artemia* assembly stages. 101
Figure 4-6. Evolution of the gap percentage in the subsequent *Artemia* assembly stages. 102
Table 4-4. Proteins found on contaminant scaffolds removed from the assembly. .. 103
Figure 4-7. Coverage distribution for the assembled *Artemia* draft genome 103
Table 4-5. Results from BLAST of 16 *Artemia* microsatellites to the *Artemia* genome assembly. 104
Figure 4-8. Occurrence of exons per gene in the *Artemia* genome 105
Figure 4-9. Completeness of the predicted coding genes 105
Figure 4-10. Components of the *Artemia* genome 106
Figure 4-11. Number of repeat library elements with a certain length. 107
Figure 4-12. Influence of repeats on genome assembly structure 109
Table 4-6. Genome annotation characteristics for *A. franciscana* and arthropod models *D. pulex*; *D. melanogaster* and *T. urticae* 111

List of abbreviations

AFLP: amplified fragment length polymorphism
AGP: a golden path
ARC: Laboratory of Aquaculture & *Artemia* Reference Center
blastn: nucleotide-nucleotide BLAST
blastp: protein-protein BLAST
blastx: translated nucleotide-protein database BLAST
bp: base pairs
BSA: bulked segregant analysis
C1: cross 1
C2: cross 2
C3: cross 3
cDNA: complementary DNA
CDS: coding DNA sequence
cM: centiMorgan
DAP: 4',6-diamidino-2-phenylindole
DNA: Deoxyribonucleic acid
ERVL: endogenous retrovirus-related LINE
ESD: environmental sex determination
EST: expressed sequence tag
Evalue: expectation value
FDR: false discovery rate
GFP: green fluorescent protein
GH: growth hormone
GIGA: global invertebrate genomics alliance
GS: genome size
GSD: genotypic sex determination
I5K: 5,000 Insect Genome Project
IAG: androgenic gland-specific insulin-like peptide
ID: BLAST identity
JH: juvenile hormone
LG: linkage group
LINE: long interspersed nuclear element
LOD: logarithm of the odds
LTR: long terminal repeat
M: million
MP: mate-pair
mRNA: messenger ribonucleic acid

mtDNA: mitochondrial DNA

N50: weighted median statistic, 50% of the bases in the assembly is contained in contigs or scaffolds equal to or larger than this value

NCBI: National Center for Biotechnology Information

NGS: next-generation sequencing

PAR: pseudo-autosomal region

PC: primer combination

PCR: polymerase chain reaction

PE: paired-end

pg: picograms

PI: propidium iodide

QTL: quantitative trait locus

RAPD: random amplified polymorphic DNA

RFLP: restriction fragment length polymorphism

RNA: Ribonucleic acid

RNAi: RNA interference

RNAseq: RNA sequencing

SFB: San Francisco Bay *Artemia* strain

SINE: short interspersed nuclear element

SNP: single-nucleotide polymorphism

tblastn: protein-translated nucleotide database BLAST

VC: Vinh Chau *Artemia* strain

WGS: whole-genome sequencing

Chapter 1

Introduction

In this chapter, we focus on *Artemia*, its taxonomy, habitat, life cycle, and unique extremophilic features and its relevant applications in the aquaculture industry and in research. Then, the status and availability of arthropod genomes is discussed and to what degree genomic research has been conducted in *Artemia* thus far. Next, sex determination in arthropods is reviewed and techniques used in the following chapters, such as linkage mapping, bulk segregant analysis and genome sequencing are explained. Finally, the aims of our research are clarified and the outline of the next chapters is provided.

1.1. *Artemia*

1.1.1. Taxonomy

Artemia, also known as brine shrimp, is an arthropod of the Pancrustacea (Figure 1-1). Pancrustacea comprise the former Crustacea as well as the terrestrial insects (such as Hexapoda), meaning that insects and crustaceans are not sister groups as supposed before, but that insects are actually a type of crustacean [169]. The Pancrustacea are divided into two clades: (1) Malacostraca, Copepoda & Cirripedia and (2) Branchiopoda, Remipedia & Hexapoda. The Branchiopoda (e.g. *Artemia* and *Daphnia*) and the Remipedia are clearly more closely related to the Hexapoda (such as insects) than to any other Pancrustacea [220]. Thus, Branchiopoda and Remipedia are interesting organisms to study the crustacean-insect branching.

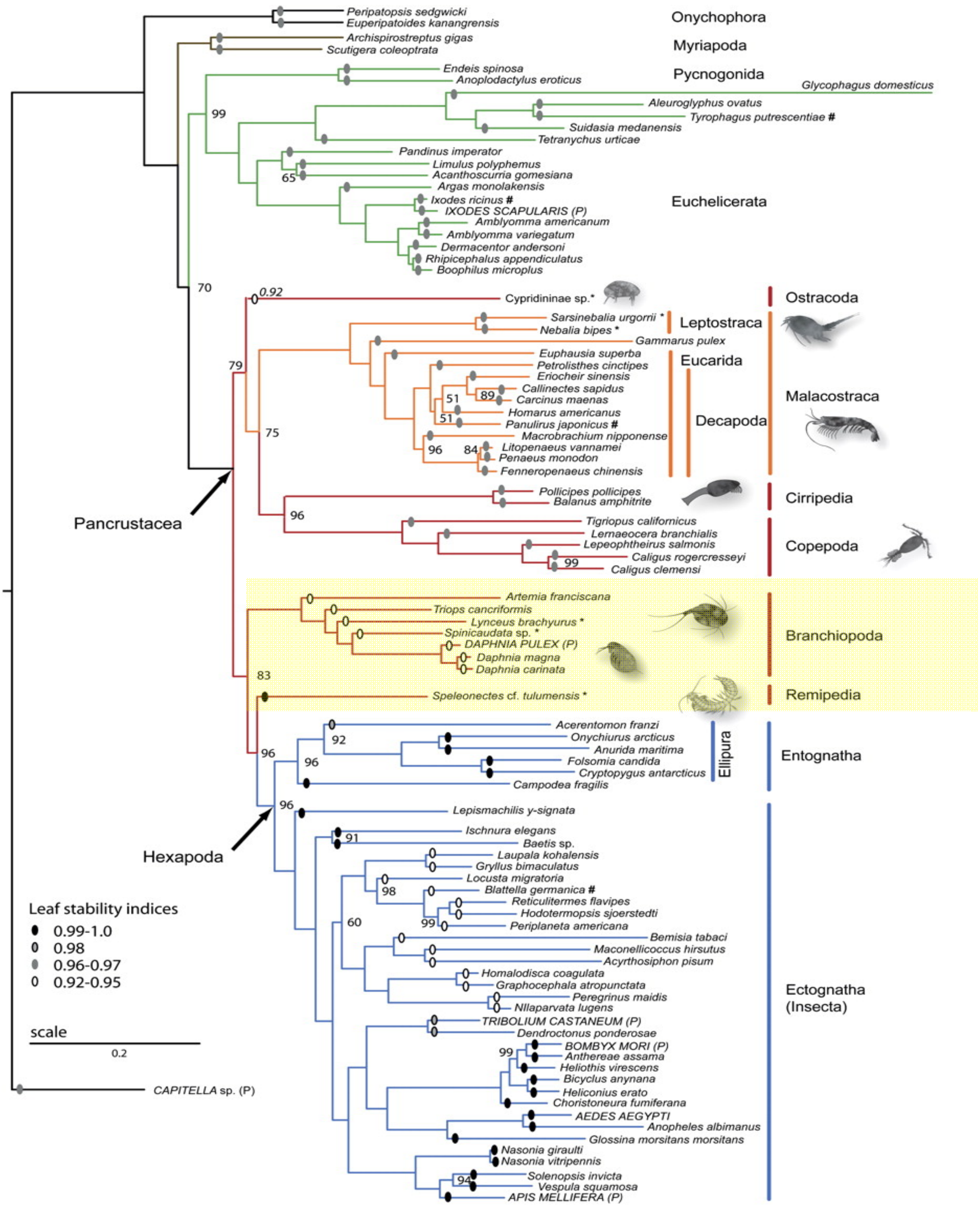


Figure 1-1. Pancrustacean phylogeny, adapted from Von Reumont et al. (2012) [220]. Branch color code: crustaceans (red and orange), hexapods (blue), chelicerates (green), myriapods (brown), and outgroup taxa (black). Branchiopoda and Remipedia are highlighted in yellow.

Artemia, as well as the water flea *Daphnia* are part of the Branchiopoda class, of which gill-like appendages are a typical feature. An overview of *Artemia* taxonomy is provided in Table 1-1.

Table 1-1. Taxonomy of the brine shrimp *Artemia* [169].

kingdom	Animalia
phylum	Arthropoda
subphylum	Pancrustacea
class	Branchiopoda
order	Anostraca (fairy shrimp)
family	Artemiidae
genus	*Artemia*

Artemia is a genus of small planktonic crustaceans found in hyper saline environments worldwide [109]. It comprises six sexually reproducing, diploid species (*Artemia franciscana, A. persimilis, A. salina, A. sinica, A. tibetiana, A. urmiana* [12]) and several obligate parthenogenetic *Artemia* populations consisting of different clones and ploidies (2n→5n) that cannot be grouped under one species [131]. Parthenogenesis is said to have arisen independently four times in the *Artemia* genus from the sexually reproducing species *A. urmiana*, *A. tibetiana* and *A. sinica*, with a first appearance at least 3 million years ago (Figure 1-2) [16]. *Artemia* species share a common ancestry in the middle Cretaceous, 90-80 million years ago according to Baxevanis *et al*. (2006).

The Laboratory of Aquaculture & *Artemia* Reference Center (ARC) at Ghent University possesses a cyst bank containing over 1,700 *Artemia* population samples collected from different locations around the world.

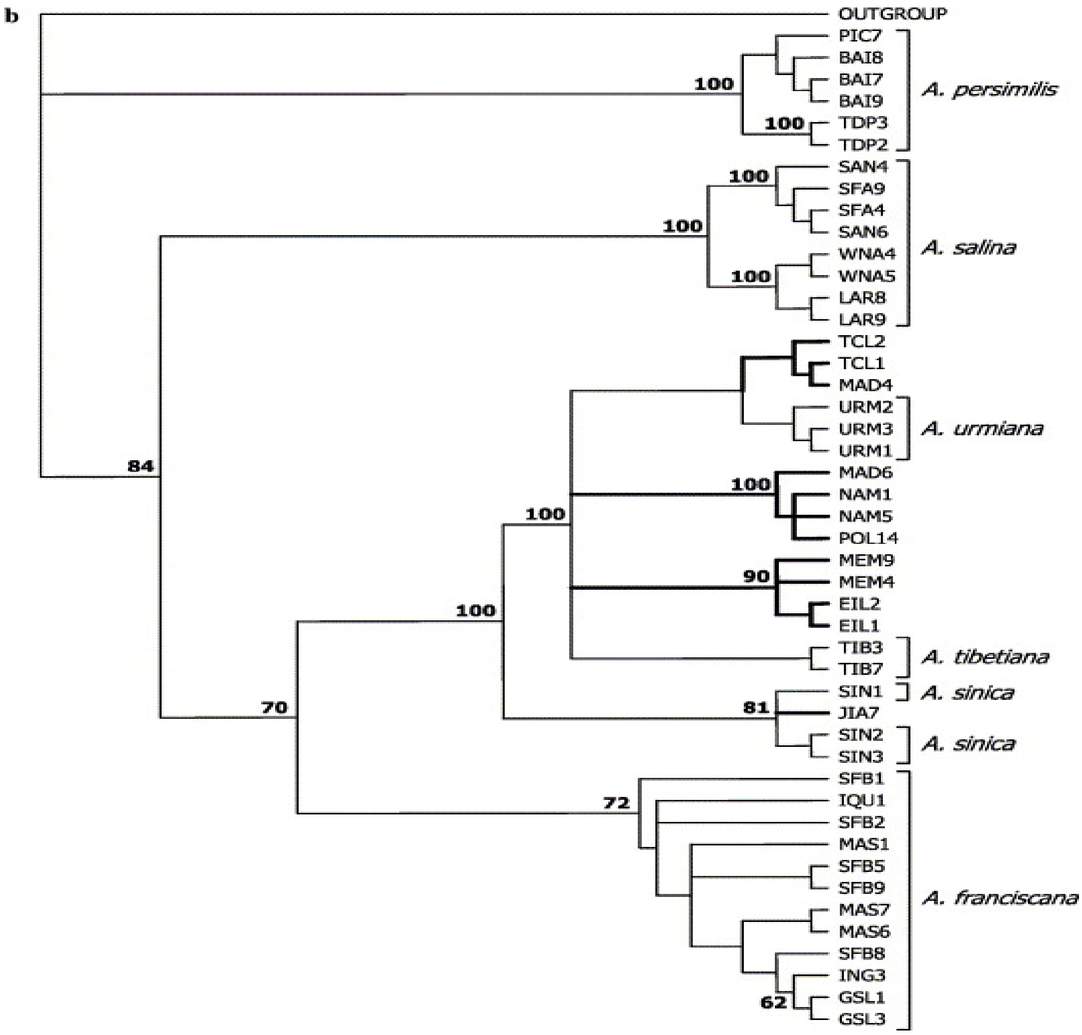


Figure 1-2. Phylogenetic relationships among *Artemia* species. Maximum parsimony tree derived from the ITS1 sequence data. Nodal support is indicated above branches. Thickened lines indicate parthenogenetic lineages. The topology is rooted with the outgroup *Streptocephalus proboscideus*. Adapted from Baxevanis *et al*. (2006) [16].

1.1.2. Habitat

Artemia has a broad geographical distribution: it is found in natural hyper saline lakes and salterns around the world. Its habitat varies regarding salinity, climate, and altitude. *Artemia* inhabits hyper saline environments of which the anionic composition can differ considerably (chlorides, sulfates, or carbonates, or combinations of up to all three [205]). The species can be found under climatological conditions that range from humid-subhumid to arid [205] and at altitudes from sea level to over 5000 m [233].

A. franciscana, originally from the American continent, has been introduced through either deliberate or inadvertent releases into Asia, Africa, Europe and Australia. It is the most studied *Artemia* species, is considered a “super species” and exhibits high levels of phenotypic plasticity [24].

1.1.3. Life cycle and extremophilic features

Many of the interesting features of *Artemia* are related to its unusual life cycle of 2 to 4 weeks (Figure 1-3).

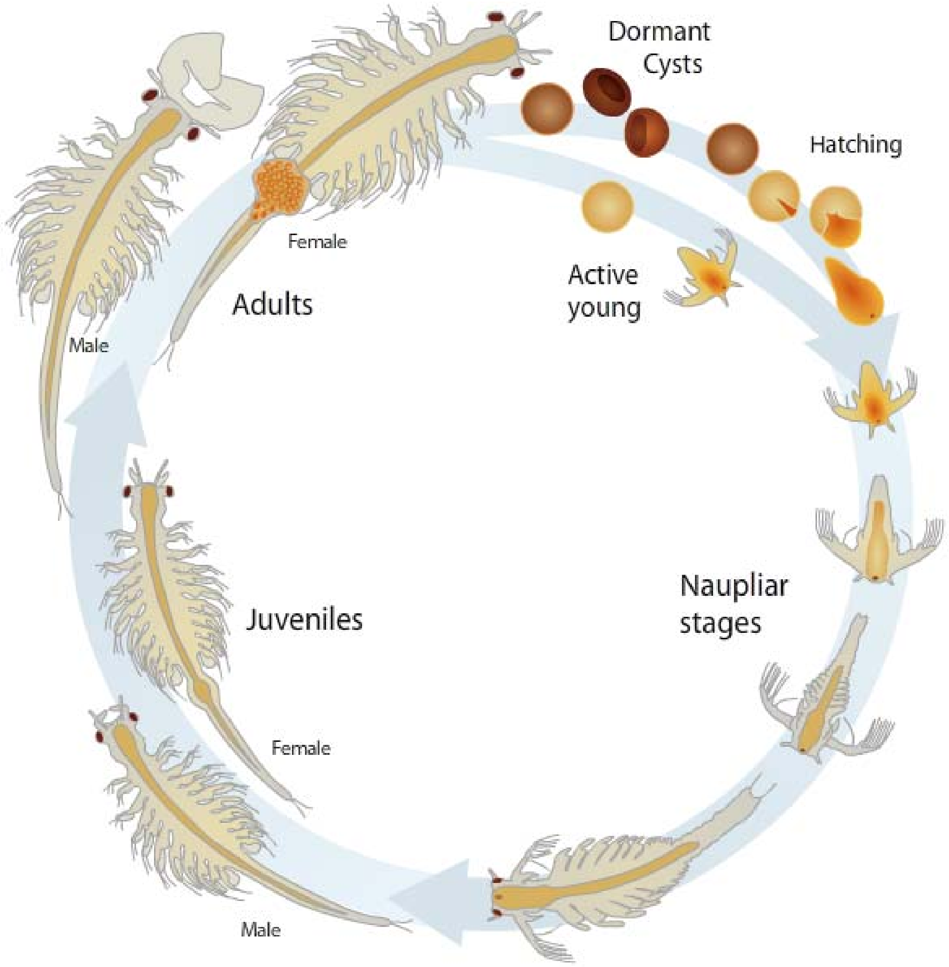


Figure 1-3. The life cycle of sexually reproducing *Artemia* [29]. Either cysts or free-swimming larvae (nauplii) are released by the female. They develop into juveniles and eventually adults. Adult males and females are distinguishable by the naked eye: males have two claspers on the head and females have an abdominal ovisac.

Under optimal conditions, adult *Artemia* females produce active free-swimming larvae (ovoviviparity), whereas under stressful conditions (such as high salinity or low oxygen levels), females produce encysted gastrula embryos or cysts (oviparity) that can remain viable for years under extreme conditions, similar to plant seeds. These cysts are in a diapaused state, during which cell division and embryonic development are fully arrested [238]. *Artemia* diapaused cysts show a 100% drop in respiration rate, the most profound metabolic arrest ever reported during diapause [36]. The diapaused state can last for very long periods. Depending on the species and on their environment, diapause can be terminated by specific environmental stimuli, such as freezing, limited oxygen levels, or cyst dehydration, all leading to a quiescent state.

As soon as 12 hours after hydration of quiescent cysts, hatching starts, and free-swimming larvae are produced.

Artemia species are extremophiles, i.e., they survive in salinity up to 10-fold that of standard seawater [66]. Osmoregulation is a fundamental aspect of physiology which is particularly important for the halophilic brine shrimp. *Artemia* can take up water by swallowing salt water and excreting the excess NaCl. Two organs are responsible for salt excretion in *Artemia* at their respective times of development: the naupliar salt gland and the juvenile thoracic epipod [143]. Besides extreme salinities, brine shrimps can withstand low atmospheric pressure, low oxygen levels and cold temperatures.

Encysted *Artemia* embryos (cysts) are probably the most stress-resistant of all animal life-history stages. They have the ability to tolerate anoxia for periods of years, while fully hydrated and at physiological temperatures. The overall metabolism of anoxic embryos is brought to a reversible standstill, including the transduction of free energy and the turnover of macromolecules. Such an extraordinary stability is partly achieved by massive amounts of a small heat shock protein (p26) that acts as a molecular chaperone [38]. Encysted embryos also have a unique tolerance for high doses of UV and ionizing radiation, thermal extremes, and desiccation-hydration cycles, more than any other desiccation-tolerant animal system [37,78,172]. The dessication tolerance in *Artemia franciscana* is attributed to the presence of biological glasses which can be altered by natural selection and thermal adaptation [78].

1.1.4. Applications in the aquaculture industry and in research

Aquaculture is the fastest-growing food production sector in the world, producing 46% of the worldwide human fish consumption [55]. *Artemia* is the most commonly used live food in aquaculture activities, specifically for larval growth of more than 85% of the marine species reared in aquaculture [98]. Annually, over 2,500 metric tons of dry *Artemia* cysts are marketed worldwide for on-site hatching to be used as fish and shellfish food. Dry cysts can be kept for years and hatched “on demand” within 24 hours to produce live food. The nutritional quality of the *Artemia* can be tailored to the larvicultural needs of many species of fish and shellfish by bio-encapsulating nutrients in the brine shrimp larvae. Bio-encapsulation of *Artemia* is defined as the enrichment of *Artemia* with specific components using their non-selective filter feeding behavior and feeding the enriched *Artemia* to the reared organism. The same bio-encapsulation method is now being developed for oral delivery of vitamins, chemotherapeutics and vaccines [109]. Adult *Artemia* is used in small amounts in shrimp brood stock diets for shrimp maturation and as an alternative for fishmeal in fish feed.

Besides their use as food, *Artemia* are also sold as pets under the marketing name Sea-Monkeys.

Many areas of scientific research investigate *Artemia*, also known as “the aquatic *Drosophila”*, mainly because of its short life cycle and its cheap, easy breeding under laboratory conditions [64]. Brine shrimps are widely used as a model organism in toxicity assays for a range of compounds such as pharmaceuticals, insecticides, metals, products of plant, fungal or bacterial origin and recently, nanoparticles, because cytotoxicity of these compounds in *Artemia* and in humans is often highly correlated [61,140,166]. *Artemia* is an efficient biofilter for bioremediation of aquaculture wastewater as well [136] and recently, has shown potential in a mitigation strategy against harmful algal blooms in coastal waters [134].

Brine shrimp species are used as a model in other research fields, such as crustacean metabolism, biodiversity studies, *Vibrio* pathogen resistance in shrimps and interactions between sexual and parthenogenetic populations [7,75,82,102,145]. They are also part of a model marine four-level food chain (algae, *Artemia*, shrimp and fish) and thus, are suitable to study trophic transfer of compounds from one marine food chain level to the next [191].

Furthermore, *Artemia* is interesting for studying evolutionary processes, because many factors thought to be responsible for genetic differentiation and speciation are observed in brine shrimp, namely: ecological isolation, formation of Dines[1] in heterochromatin content, poly-, hetero- and aneuploidy, parthenogenesis and pre- or post-mating reproductive isolation [161]. *Artemia* has proven to be an invaluable genetic model for fine-scale studies of micro evolutionary divergence, as shown by the significant divergence of *A. franciscana* strains, already one year after introduction into Vietnam [95]. Brine shrimps are phylogenetically close to the insect-crustacean branching (Figure 1-1), making them suitable to better understand the origin and evolution of insects from a crustacean ancestor.

Genes underlying the extreme phenotypes of *Artemia* might be of utmost interest, and make *Artemia* a promising model for stress response studies. Several interesting genes related to the extremophilic nature of *Artemia* have already been discovered, such as the osmoregulation gene *anterior pharynx-defective 1* [31], the cell cycle arrest termination gene *ribosomal s6 kinase* with direct applications in cancer treatment research [45] and genes coding for *Artemia* Late Embryogenesis Abundant (LEA) proteins, enhancing desiccation tolerance in mammalian cells, thus enabling engineering of biostable dried cells [116].

---

[1] interspersed elements from a transposition, over 5 million years old, well-known in *Drosophila*

1.2. Genomics

1.2.1. Genomics and gene discovery: main techniques used

1.2.1.1. Genome mapping

Genome mapping is used to localize genes in the genome of an organism. Knowledge of the genomic position of genes as a quantitative trait locus (QTL) enables the development of breeds with a specific set of beneficial and economically valuable alleles. Once the precise position of a gene has been discovered at the DNA level by means of positional cloning and/or reversed genetics, the function and structure of the gene can be characterized. The presence and order of genes on chromosomes often slightly differ between related species. Comparison of these differences through comparative genomics allows the investigation of the evolutionary mechanisms that cause species divergence [74].

Currently, two kinds of genome maps are used. First, the recombination-based genetic linkage map in which the recombination frequency is a relative measure for the distance (cM) between two markers or genes. Closely linked sequences are more often inherited together, hence they will have a lower recombination frequency [32,74]. The second type of genome map is the physical map or genome sequence, which gives absolute distances (bp) between markers or genes. No linkage maps or physical maps for *Artemia* are currently available.

AFLP is a multilocus DNA fingerprinting technique based on selective polymerase chain reaction (PCR) amplification of restriction fragments from a total digest of genomic DNA producing AFLP markers that can be used for genetic linkage mapping [223,225]. AFLP markers are typically generated in five steps (Figure 1-4). (1) Restriction digestion of genomic DNA is done with a combination of a rare (such as *Eco*RI) and a frequent cutter (such as *Mse*I). The frequent cutter is used to generate fragments of 50 to 500 bp, a fragment length that can be resolved by electrophoresis. The rare cutter is used to limit the number of fragments that can be amplified and, thus, defines the number of AFLP amplicons. (2) Double-stranded *Eco*RI- and *Mse*I- specific adapters are ligated to the restriction fragment ends. (3) A pre-amplification PCR step is performed with primers matching the *Eco*RI- and *Mse*I- specific adapter sequences and each carrying one selective nucleotide at their 3’-end with extension into the restriction fragments, to amplify only subsets of the restriction fragments to a detectable level, thus reducing the complexity of the template mixture. (4) In a selective PCR-amplification step, additional selective nucleotides are added to the *Eco*RI and *Mse*I primers to amplify subsets of the preamplified fragments, hence producing AFLP fragments.

AFLP fingerprinting of relatively large genomes (> 100 Mb) is usually carried out with two or three selective bases in one or both primers. For detection of the AFLP fragments, a radioactive or fluorescent label is added to either the *Eco*RI or the *Mse*I primer. (5) The AFLP fragments are fractionated per size by electrophoresis and displayed on denaturing polyacrylamide gels. In conventional gel electrophoresis with radiolabeled primers, gels are either dried on paper or fixed on glass plates after electrophoresis. AFLP images may be generated either by conventional autoradiography or phosphorimaging technology. When gel electrophoresis with infrared dye (IRD) or fluorescently labeled primers is applied, AFLP images may be generated with LI-COR.

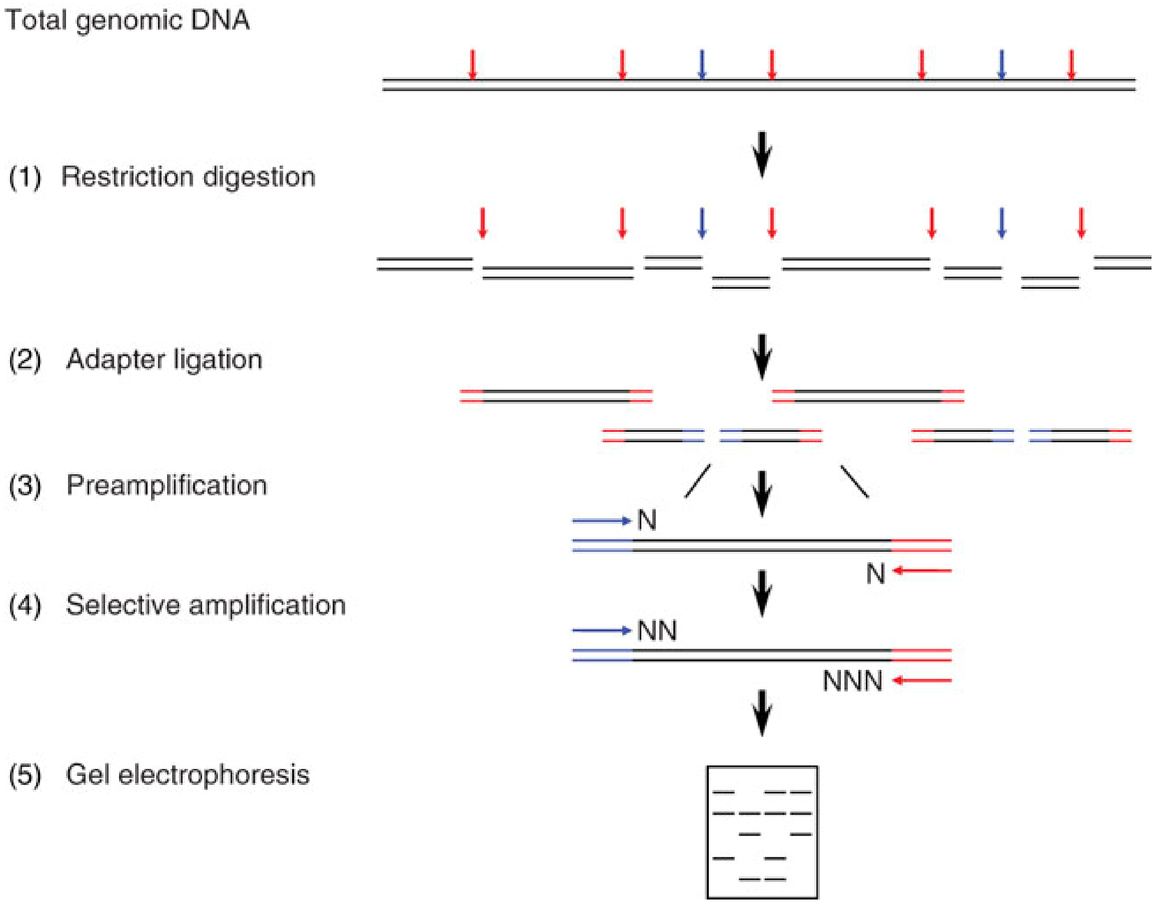


Figure 1-4. Outline of the AFLP procedure [225]. (1) Restriction digestion of total genomic DNA with rare cutter *Eco*RI (blue arrow) and frequent cutter *Mse*I (red arrow); (2) Adapter ligation of *Eco*RI- (blue) and *Mse*I- (red) specific adapters to the fragment ends; (3) PCR preamplification of subsets of the *Eco*RI/*Mse*I templates with primers matching the adapter sequences and each carrying a selective nucleotide (represented by N); (4) Selective amplification by PCR in which additional selective nucleotides are added to the *Eco*RI and *Mse*I primers; and (5) Gel electrophoresis of the *Eco*RI/*Mse*I amplification products.

The advantages of the AFLP technique over other formerly used DNA marker technologies are its cost efficiency, higher reproducibility than RAPD, that prior sequence information from the study organism such as a physical map is not required and it generates many polymorphic markers in a single PCR.

These advantages make AFLP suitable for construction of high-resolution genetic linkage maps [223], which are critical for gene identification through positional cloning and determination of the genetic basis of quantitative traits [237]. Therefore, AFLP is widely used for genetic mapping in plant and animal species. One disadvantage of AFLP markers used to be they could only be scored as dominant markers (present/absent), but techniques have been developed to score markers codominantly (intensity), allowing both dominant and co-dominant AFLP markers to be used for a more efficient and faster map construction. Scoring of AFLP markers can be done based on relative fragment intensities, using the image analysis software AFLPQuantar*Pro* (http://www.keygene-products.com) [225].

A genetic linkage map shows the genetic distance between genetic markers, defined as a function of the crossover frequency during meiosis. The closer two markers are on a chromosome, the lower the chances are that crossing over will occur between them. Thus, how often those two markers are inherited together in a population after sexual reproduction reflects their linkage. Linkage analysis of both dominant and codominant markers is performed with mapping software, such as Joinmap [211], a computer program for the calculation of genetic linkage maps in experimental populations of diploid species. Markers are encoded and linkage groups are determined based upon four criteria: (1) the independence test logarithm of odds (LOD) score, (2) the linkage LOD score, (3) the independence test P-value and (4) the recombination frequency. Once the linkage groups are determined, the linkage map can be calculated for each group. Only linkage groups containing at least three markers are meaningful for map construction. Maps are constructed in three rounds, each producing a linkage map. Each map is calculated by using the pairwise data of loci present on the map. Once the well-fitting markers (causing a change in goodness of fit smaller than the threshold = 5) are positioned on the map (after two rounds), the remaining markers are forced onto the map by setting the jump threshold to zero. When the markers in the third map cause a jump in goodness of fit larger than an arbitrary threshold, the second map is selected as the final map instead of the third map. A marker order is not forced on any linkage group during map construction.

1.2.1.2. Bulked segregant analysis

Quantitative or complex phenotypes are traits controlled by multiple genes and environmental factors. A primary challenge in genetic research is the identification of the genome parts that contribute to variation in complex traits (QTL), and, ultimately, of the genes and alleles responsible for trait variation. Usually, genotyping with molecular markers of each individual from a segregating population is required to locate QTL.

However, individuals from such a segregating population can also be bulked according to phenotype and investigated for presence/absence of marker alleles (e.g. AFLP) or for differences in allele frequencies with high-throughput sequencing to produce short sequence reads [128]. Such an approach is designated as bulked segregant analysis (BSA) and is very helpful in mapping traits quickly, efficiently and relatively inexpensively in both model and non-model organisms [141,165]. The success of the BSA approach relies on the considerable reduction in the amount of marker assays compared to building a genetic map for identification of genetic markers associated with a phenotype in each individual of the mapping population.

Highly multiplexed markers such as single-nucleotide polymorphisms (SNPs) have the potential to further reduce the number of marker assays into a single whole-genome assay, resulting in so-called “fast-forward genetics”. A statistical framework has been developed for QTL mapping strategies with BSA and next-generation sequencing (NGS) techniques, suggesting that with large bulks and sufficient sequencing depth, these strategies can be used to detect even weak-effect QTLs [128]. This concept has been proven by application in the budding yeast *Saccharomyces cerevisiae* [128]. Limitations of BSA are that it can only be used for studying one single common background and cannot be interpreted reliably when epistatic interactions perturb the studied traits [83]

With a combined approach of BSA and NGS, NGS reads need to be produced, a reference genome needs to be available or assembled with the available sequences and the NGS reads of the respective bulks need to be mapped to the genome to identify SNPs. The Burrows-Wheeler Alignment tool (BWA) efficiently aligns short sequencing reads against a large reference sequence, allowing mismatches and gaps [113]. BWA supports reads from the Illumina platform and outputs alignment in the standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses after the alignment can be achieved with the open source SAMtools software package [114].

Most of the published variant callers for next-generation sequencing data employ a probabilistic framework, such as Bayesian statistics, to detect variants and assess confidence in them. These approaches can be confounded by numerous factors such as extreme read depth and pooled samples. In contrast, VarScan employs a robust heuristic/statistic approach to call variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance. VarScan is a platform-independent software tool to detect variants in NGS data from individual or pooled samples [105].

1.2.1.3. From organism to genome

Once DNA has been extracted from the studied organism, three steps are presently needed to produce a genome sequence: sequencing, assembly, and annotation.

-a) Sequencing

DNA sequencing

DNA sequencing is any method or technology, used to determine the exact order (sequence) of nucleotides (adenine, guanine, cytosine, and thymine) within a DNA molecule. Whole-genome sequencing provides read data for the complete DNA sequence of an organism's genome, including chromosomal DNA, but also mitochondrial DNA and, for plants, chloroplast DNA. Methodologies evolve rapidly because sequencing technologies improve and more genome sequences become available.

Sanger sequencing was developed by Frederick Sanger (1975) and started off as the golden standard for genome sequencing. Sanger sequencing is based on the selective incorporation of chain-terminating dideoxynucleotides that cause DNA polymerase to cease DNA extension during *in vitro* DNA replication [181]. The dideoxynucleotides are each radioactively or fluorescently labeled for detection in automated sequencing machines. The highly cited *Drosophila* [6] and *Homo sapiens* [213] genomes were sequenced using this platform. Currently, Sanger sequencing has become more time-consuming and expensive than newer and improved technologies. Sanger sequencing is now largely replaced by NGS, especially for large-scale, automated genome analyses. The Sanger method still remains in wide use for smaller-scale projects and enables production of long reads (average length 500-800 bp).

Sequencing platforms

NGS (or second-generation sequencing) was first introduced in 2005 (454 Life Sciences, Roche platform) and was based on pyrosequencing, namely the detection of pyrophosphate release upon nucleotide incorporation rather than on chain termination with dideoxynucleotides, as done in Sanger sequencing [154]. Since then, many NGS platforms have been developed, of which Illumina is the current market leader [135]. Illumina sequencing is based on reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands (sequencing-by-synthesis). This platform can be used for DNA sequencing as well as RNA sequencing, is very accurate and is less sensitive to homopolymers than 454 sequencing.

NGS is considered a cheap, high-throughput method: it produces more than 100-fold more data per run than Sanger sequencing [154], resulting in large datasets of short reads (often 50-100 bases for Illumina, 1000 bases for 454 platforms). The large amounts of reads make NGS ideal for SNP studies. A limitation of both the 454 and the Illumina platforms is the amplification step used before sequencing, causing systematic errors. Genome areas with a high GC content or repeats are often not amplified and thus not sequenced. The PCR amplification of GC-rich DNA is often problematic due to stable secondary structures in the DNA that are resistant to melting [123]. These secondary structures cause DNA polymerases to stall, resulting in incomplete and non-specific amplification. Also, the use of short reads breaks the original structure and haplotypes of the genome, especially in repeat-rich genome areas.

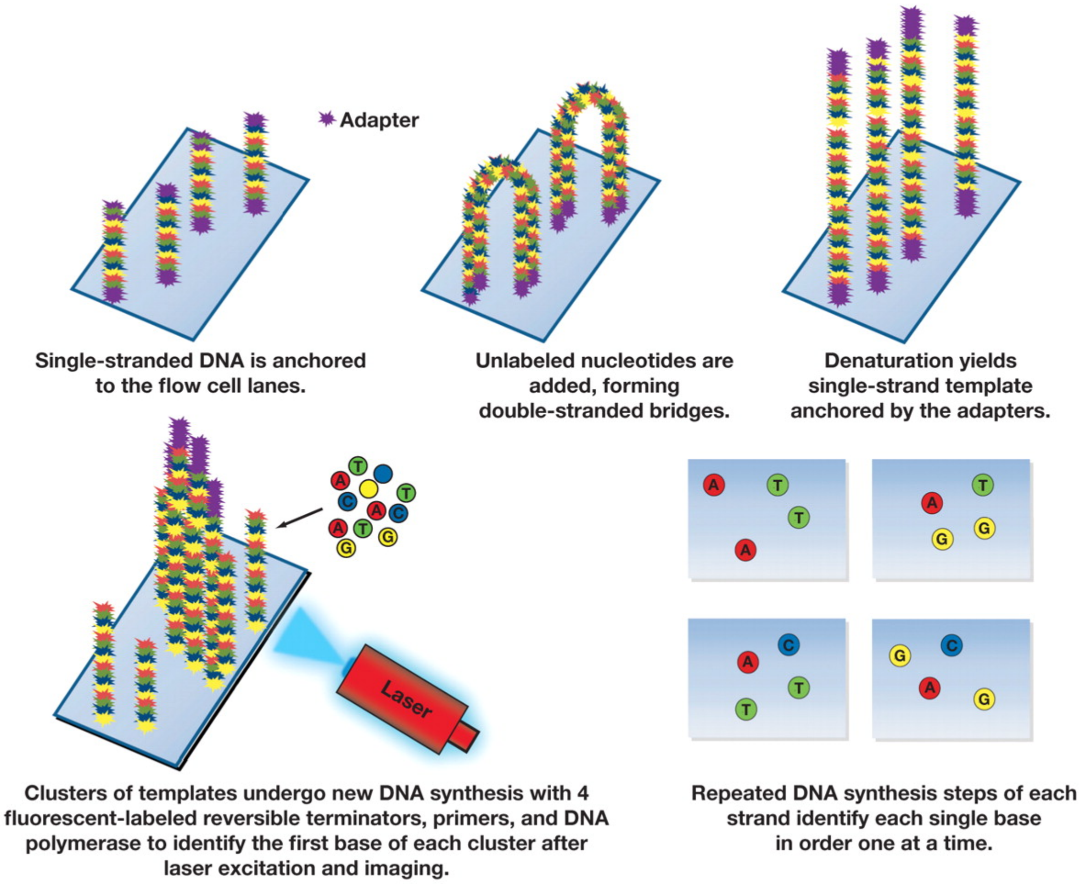
The last trends in NGS include techniques to lengthen reads [159], striving for Sanger-like read lengths, while still benefiting from NGS scale and prices (e.g. pseudo-Sanger sequencing). Another trend is the enlargement of insert sizes, mainly for an improved assembly result of repeat-containing regions [209].

An ideal sequencing strategy would be to determine the entire chromosome sequence in one long contiguous read instead of chopping the genome in small pieces and then assembling it back together with bioinformatics software, as is currently the practice when using NGS. Very recently, third-generation sequencing has arisen, following the concept of sequencing-by-synthesis, but without a prior amplification step, i.e., single-molecule sequencing, with Pacific Biosciences SMRT [154]. Benefits of Pacific Biosciences are that no amplification step is needed, so no GC bias is present, reads are long (average length of 7 kb), so many repeats are sequenced, haplotypes can be better resolved, SNP haplotypes can be inferred, and very accurate assemblies are created (e.g. *Arabidopsis* [152]). The most important limitations of the Pacific Biosciences platform are currently the prohibitive prices for large genomes, the high standards needed for DNA extraction and library preparation as well as the high random sequencing error rate, although the effect of the latter can be countered by using a minimal coverage of 10X and adequate data processing.

Hereafter, NGS will refer to the Illumina platform only.

Illumina sequencing: how it works

First, the library is prepared: DNA fragments of the desired size (insert size) are produced and amplified and adapters are ligated to both ends of the DNA fragments. Then, the libraries are loaded into the Illumina sequencer and sequencing starts (Figure 1-5).



Single-stranded DNA is anchored to the flow cell lanes.

Unlabeled nucleotides are added, forming double-stranded bridges.

Denaturation yields single-strand template anchored by the adapters.

Clusters of templates undergo new DNA synthesis with 4 fluorescent-labeled reversible terminators, primers, and DNA polymerase to identify the first base of each cluster after laser excitation and imaging.

Repeated DNA synthesis steps of each strand identify each single base in order one at a time.

Figure 1-5. Illumina sequencing scheme by Ross & Cronin [174]. From left to right and from top to bottom: 1) Binding of single-stranded, adapter-ligated fragments to the flow cell surface; 2) Bridging of a ligated fragment to a complementary oligonucleotide on the flow cell surface; 3) Denaturation of bridges to one-stranded fragments; 4) Localized amplification of single molecules leading to DNA cluster formation, amplification of clusters using fluorescently reversible termination bases, laser removal of termination bases, imaging; 5) Repetition of DNA synthesis for each nucleotide of the DNA molecule until it is completely sequenced.

The flow cell surface of the sequencer is coated with a dense layer of single-stranded primers, to which the adapters on the denatured DNA bind. In contrast to the Roche method, in which a bead-based emulsion PCR is used to generate “colonies”, Illumina utilizes a unique "bridged" amplification reaction that takes place on the flow cell surface. Single-stranded, adapter-ligated fragments bound to the flow cell surface are exposed to reagents for polymerase-based extension. Priming happens as the free/distal end of a ligated fragment "bridges" to a complementary oligonucleotide on the surface. The bridges become double-stranded and are denatured to one-stranded fragments, only attached to the flow cell by one end. These one-stranded fragments become double-stranded again, completing the first amplification. Repeated denaturation and extension result in localized amplification of single molecules (“DNA clusters”) in millions of unique locations across the flow cell surface.

The four types of reversible termination bases (adenine, cytosine, guanine, and thymine), each fluorescently labeled with a different color and attached to a blocking group, are added. The four bases compete for binding sites on the template DNA to be sequenced and non-incorporated molecules are washed away. After each cycle, a laser is applied and a detectable fluorescent color specific to one of the four bases is then visible for each cluster, allowing sequence identification. The 3'-terminal blocking group and the probe are then removed and the sequencing of the next base in the DNA molecule starts. The process is repeated until the full DNA molecule is sequenced.

Paired reads, such as paired-end reads and mate-pair reads are mainly used for more accurate mapping of new DNA or RNA reads to the reference genome of an organism, for scaffolding of genomes (see section 1.2.1.3. -b) or for assembling long genomic regions in the presence of repeats. The larger the insert size, the longer the repeats that can be bridged by paired reads. Paired-end and mate-pair Illumina technologies rely on the sequencing of both ends of a sequence fragment (insert). In paired-end sequencing, genomic DNA is broken into fragments of 200-500 bp and adaptors are ligated to both ends of the fragments and bind to the flow cell, forming clusters. One end of the fragment is sequenced, the clusters are regenerated and the other (paired) end is sequenced. Mate-pair sequencing follows a similar procedure, but because it is used for longer inserts (3-5kb), the fragments first need to be blunt-end circularized by biotinylation, fragmented again and enriched for biotinylated fragments.

Mate-pair (MP) reads that pass through the junction of the two joined ends of a 3 to 5 kb DNA fragment are not easy to identify and cause problems during mapping and *de novo* assembly. The risk for this situation increases when longer read lengths are implemented. To solve this problem, a protocol for use with Illumina sequencing technology was developed, using Cre-Lox recombination in which a LoxP sequence (a tag) is incorporated at the junction site instead of the classic blunt-end circularization [210]. This tag allows screening of reads for junctions without a reference genome. Identified junction reads can be trimmed or split at the junction. Moreover, the location of the tag in the reads distinguishes mate-paired reads from spurious paired-end reads. These are the reasons why cre-lox MP sequencing improves the assembly result compared to tagless mate-pairs [210].

-b) Assembly

Trimming

Before assembling genome reads from an NGS platform, low-quality bases, more likely to contain sequencing errors, primers and adapters from the sequencing process, are best removed first with quality and primer/adapter trimming software [99]. This improves assembly errors and gaps in the subsequent assembly and enhances variant detection. Alignment of variable regions can show mismatches due to both polymorphisms and sequencing errors, hence the improvement of sensitivity for variant detection upon removal of sequencing errors [99].

De novo assembly of a genome

The choice of the assembly software depends on the genome size and structure (repeat content), the amount of read data, the available computing power and time, and the read length, which, in turn, depends on the chosen sequencing platform. In the *de novo* assembly approach, without the use of a reference genome, sequence reads are compared to each other, and then overlapped to construct longer contiguous sequences. Basic steps are shown in Figure 1-6.

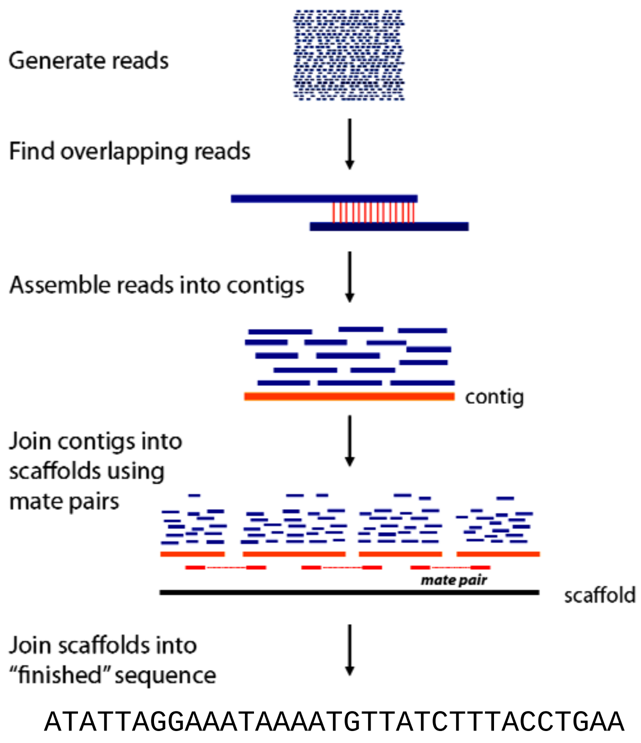

Generate reads
Find overlapping reads
Assemble reads into contigs
contig
Join contigs into scaffolds using mate pairs
mate pair
scaffold
Join scaffolds into "finished" sequence
ATATTAGGAAATAAAATGTTATCTTTACCTGAA

Figure 1-6. Basic steps in sequence assembly, scaffolding and genome finishing [151]. Reads are assembled into contigs with the de Bruijn graph approach or the overlap/layout/consensus approach. Contigs are joined into scaffolds using paired information from mate-pair and/or paired-end reads. Scaffolds are joined into chromosome-length fragments by genome finishing techniques.

Assembly of a large genome with only short Illumina reads requires using a computer that has a very large amount of RAM, a computer cluster or expensive commercial software. For the *de novo* assembly of the large *Artemia* genome at the starting time of this project, several potential assembly programs known to assemble large genomes were considered: SOAPdenovo, ABySS, Celera wgs Assembler and CLC Assembly Cell.

During the assembly from reads to contigs, standard assembly algorithms are used: the overlap/layout/consensus approach (as in Celera wgs Assembler), the Eulerian path approach (as in ABySS and SOAPdenovo) and the de Bruijn graph approach, which is a fast and computationally efficient combination of both previous approaches (as in CLC Assembly Cell) [142].

SOAPdenovo is an all-purpose genome assembler which was used to assemble the giant panda genome [115]. Benefits of SOAPdenovo are that it is open-source (free) software, relatively fast, and it contains a scaffolder and a gap-filler. Disadvantages are the large amount of RAM needed and the relative greediness of the assembly algorithm, causing long, but often incorrect genome fragments to be produced [179]. ABySS is a *de novo* short read assembler which can run on multiple nodes using a computer cluster, thus the amount of RAM needed per cluster node is relatively smaller [188]. A disadvantage is that it is relatively slow. Celera wgs assembler was used to produce the first human genome assembly [213]. Benefits of Celera are that it is open-source software and it can assemble data from most common sequencing platforms. Disadvantages are the amount of time and RAM needed for assembly. CLC is a commercial genome assembler which is very fast, uses little RAM, can run on a computer cluster, contains a scaffolder and can assemble data from most common sequencing platforms [35]. The disadvantage is that it is not open-source.

Scaffolding and gap filling

After reads are assembled into contigs, the contigs can be joined in a scaffolding step by using the paired information from paired-end and mate-pair reads that span two contigs to assess the order, distance and orientation of contigs and combine them into scaffolds [19] (Figure 1-6). SSPACE is a stand-alone scaffolder of pre-assembled contigs using paired-read data. Main features are: a short runtime, multiple library input of paired-end and/or mate pair datasets and possible contig extension with unmapped sequence reads. SSPACE shows good results on both prokaryote and eukaryote genomic test sets, where the amount of initial contigs was reduced by at least 75% [19].

Where the contigs are joined together into scaffolds, unknown bases can be reduced by introducing a gap filling step, again by means of paired read information [20]. Gap filling software GapFiller gives better results than gap filling with SOAPdenovo both on bacterial and eukaryotic datasets and accurately closes most of the gaps, using only a limited amount of memory [20]. GapFiller fills gaps automatically by finding read pairs of which one member matches within a sequence region and the second member falls (partially) within the gap. The latter reads are then used to close the gap through sequence overlap. Gaps are entirely closed only if the size of the sequence insertion corresponds closely to the estimated gap size after scaffolding, which is based on the alignment of paired reads to the contigs. This process is iteratively repeated until no further gaps can be closed.

Genome finishing

The last step in genome assembly is genome finishing, in which chromosome-length fragments are produced from shorter genome fragments (usually scaffolds; Figure 1-6). Whereas high-throughput short-read sequencing such as NGS will yield high quality assemblies for small genomes such as bacteria, it is impossible to completely and accurately assemble large, complex genomes of plants and animals without other long-range contiguity information. Therefore, genome finishing requires additional sequencing of long fragments (single-molecule sequencing, Sanger sequencing or pseudo-Sanger sequencing), NGS with very large insert sizes (mate-pair sequencing), or genome mapping (optical mapping [107] or high-resolution genetic mapping) and often, a combination of these methods [194].

-c) Quality characteristics of a *de novo* genome assembly

Traditional methods to determine assembly quality are based on standard metrics such as the N50 value, the number and length of contigs in the assembly, gap% and gap length and coverage [219]. The contig or scaffold N50 is a weighted median statistic so that 50% of the nucleotides in the entire assembly are contained in contigs or scaffolds equal to or longer than the N50 value. The sequencing depth or coverage of a genome is generally expressed by the average number of times each base of the genome has been sequenced. Coverage can be based on either the estimated genome size or the genome assembly size. Standard metrics focus on size without taking into account the correctness of the assembly, which can be best evaluated from concurrence with expression data and proteins from the studied organism and from the quality of the genome annotation results.

-d) Genome Annotation

Once the assembly is considered completed, the genome can be annotated. Genome annotation is the process of linking biological knowledge to the genome sequence.

Repeats

The first step before annotation is repeat identification and masking. In the annotation community, “repeat” means “low-complexity” sequences (such as homopolymeric runs of nucleotides), as well as transposable (mobile) elements, i.e. viruses, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) [229]. Repeats can be identified with the software RepeatMasker [198].

Gene prediction

The second step is (*ab initio* and/or evidence-driven) gene prediction: identification of non-protein-coding portions and of coding elements in the genome by alignment of all available transcriptome and proteome data of the studied organism as well as those of related species.

The accurate prediction of intron boundaries largely facilitates the correct prediction of gene structure in nuclear genomes. Many tools for localizing these boundaries on DNA sequences have been developed, but these tools still make many false positive predictions. Open-source splice site prediction tool SpliceMachine shows state-of-the-art prediction performance on *Arabidopsis thaliana* and human sequences, performs a computationally fast annotation, and can be trained by the users on their own data, such as junctions from RNA-seq [46]. Training of SpliceMachine is achieved by extracting donor and acceptor information from junction data obtained from mapping RNAseq data onto the genome, typically using a positive/negative ratio of 1/10.

The platform EuGène [182] provides an integrative approach to produce high-quality automatic annotations for genomes of higher organisms. *Ab initio* training of software such as EuGène is typically done by selecting 100-200 structurally correct (manually curated) and representative genes in a genomic context of coding and non-coding sequence. Coding regions are identified with an Interpolated Markov Model (IMM) in EuGène. All annotation input (e.g. junctions from RNA-seq, ESTs, proteins) is weighed and integrated by EuGène that will iteratively compute optimal annotation parameters, based on the selected training set of genes.

Functional annotation

The third step is functional annotation, by which biological information is linked to the coding elements by synthesis of the gene prediction data into gene annotations [229]. Because of the project-specificity of each annotation process and the involvement of many different software tools, the program collection to create a genome annotation is generally referred to as "the annotation pipeline". In Figure 1-7, three basic approaches to genome annotation and some common variations are shown: (1) gene prediction only, (2) gene prediction with a choice of the best consensus model (based on CDS) for each gene, or (3) full-scale annotation pipelines containing steps for gene prediction with a choice of the best consensus model (based on CDS, mRNA or other evidence) and finally, manual curation.

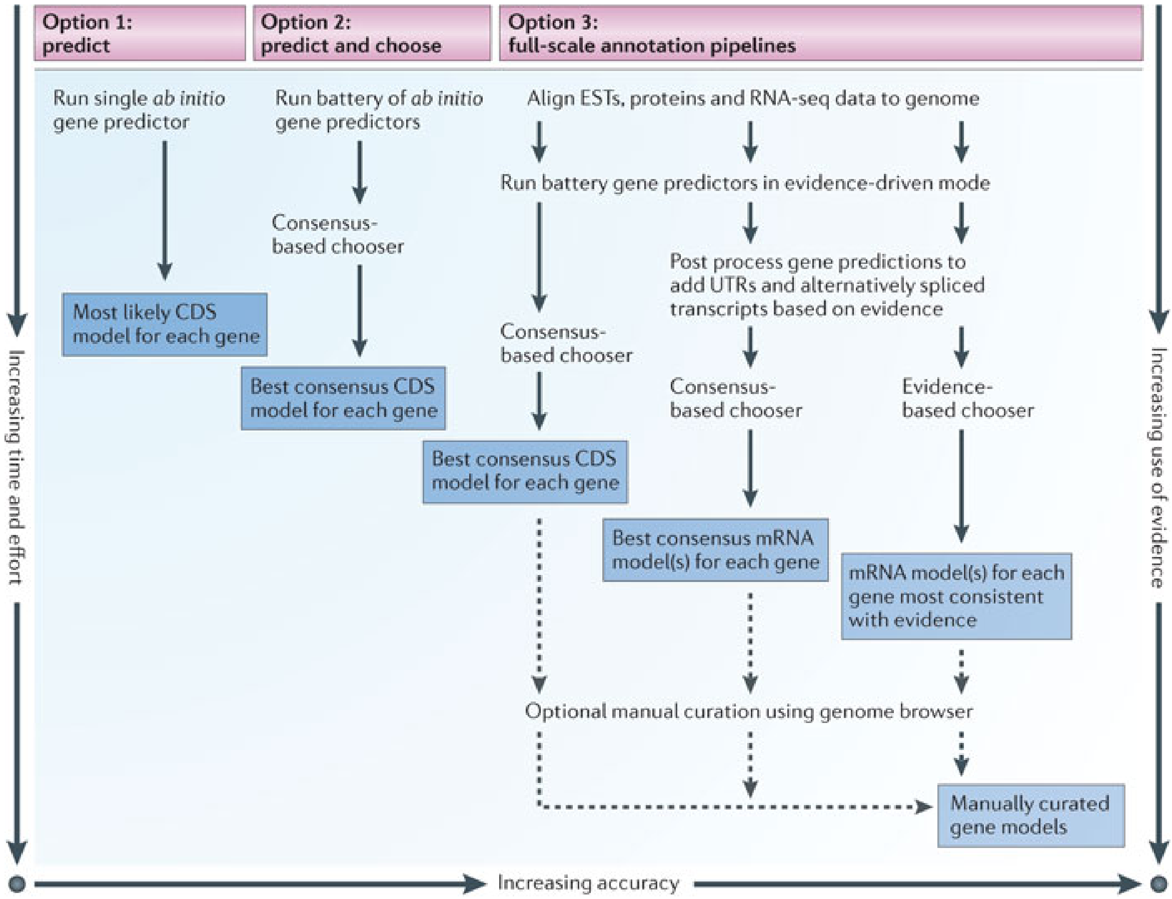

Option 1: predict
Option 2: predict and choose
Option 3: full-scale annotation pipelines
Run single ab initio gene predictor
Run battery of ab initio gene predictors
Align ESTs, proteins and RNA-seq data to genome
Run battery gene predictors in evidence-driven mode
Consensus-based chooser
Post process gene predictions to add UTRs and alternatively spliced transcripts based on evidence
Most likely CDS model for each gene
Consensus-based chooser
Best consensus CDS model for each gene
Consensus-based chooser
Evidence-based chooser
Best consensus CDS model for each gene
Best consensus mRNA model(s) for each gene
mRNA model(s) for each gene most consistent with evidence
Optional manual curation using genome browser
Manually curated gene models
Increasing time and effort
Increasing use of evidence
Increasing accuracy

Figure 1-7. Three basic approaches to genome annotation and some common variations [229]. Approaches are compared based on relative time, effort and the degree to which they rely on external evidence, as opposed to *ab initio* gene models. The y axis shows increasing time and effort; the x axis shows increasing use of external evidence and, consequently, increasing accuracy and completeness of the resulting gene models. The type of final product produced by each kind of pipeline is shown in the dark blue boxes. Relative positions in the figure are for summary purposes only and are not based on precisely computed values. CDS, coding sequence; EST, expressed sequence tag; RNAseq, RNA sequencing; UTR, untranslated region.

-e) NGS genomes: facts to keep in mind

NGS is a high-throughput technique that allows a high sequencing coverage of each base in the genome and thus, is particularly suitable for variant studies, such as SNP detection. With the assembly of the giant panda genome in 2010, *de novo* assembly of a large complex genome with NGS reads in a range of short and long insert sizes from a single individual has been proven possible [115]. Short reads nevertheless require different and much more challenging assembly strategies. They often result in genomes that are fragmented into many contigs and have missing or misassembled repeat regions [138]. This has repercussions on downstream genome analysis and requires increased annotation efforts, specifically for gene prediction at the contig ends.

1.2.2. Arthropod genome projects

Currently, the genome of 60 arthropod species has been sequenced, of which 52 insects, five arachnids (mites and ticks), one centipede and two crustaceans: the water flea *Daphnia pulex* (Branchiopoda) [39] and the salmon louse *Lepeophtheirus salmonis* (Copepoda) [2].

The sequencing of a genome is becoming ever faster and cheaper. However, data processing through bioinformatics analysis (*de novo* assembly, alignment, annotation and comparative genomics) is often a major bottleneck towards an annotated genome of the quality of, for instance, the fruit fly *Drosophila*. Many genome sequencing projects for crustaceans are still in progress, such as for the 20 crustacean species nominated for i5k, an initiative aiming at sequencing 5,000 arthropod genomes (Figure 1-8).

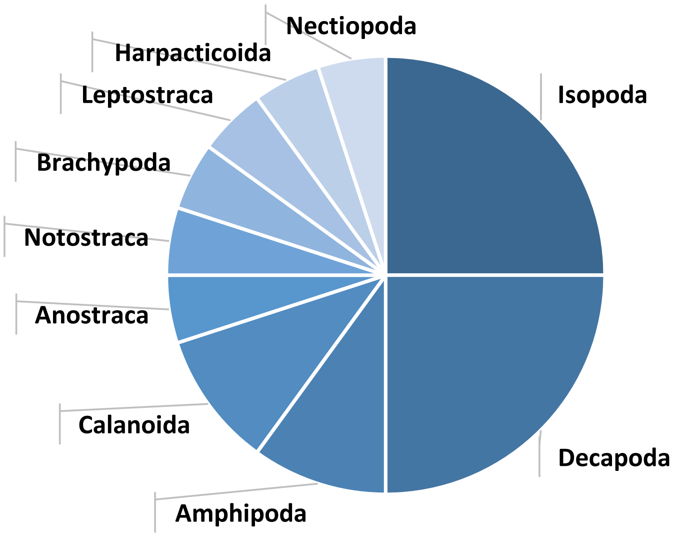


Figure 1-8. Crustacean species selected for the i5k project grouped by order, modified from the i5k nominations [1].

The Global Invertebrate Genomics Alliance (GIGA) also targets crustacean genomes, as part of their united effort to sequence the broad spectrum of non-insect and non-nematode invertebrate genomes [70].

1.2.3. *Artemia* genomics

All sexually dimorphic *Artemia* species among which *A. franciscana*, are diploids with 42 chromosomes (Figure 1-14), with the exception of *A. persimilis* (2n=44) [157]. Parthenogenetic *Artemia* populations range in ploidy from 2n to 5n and are occasionally heteroploids [131].

Artemia genome size estimations with different techniques have produced discordant estimates of 1.5 pg [212] and 3 pg [171]. In comparison with the mean crustacean genome size of 4.45 pg [73], the *Artemia* genome is relatively small.

Molecular markers used to date for population genetic, phylogeographical and evolutionary studies of *Artemia* include allozyme, random amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), AFLP markers and microsatellites [144,197,204]. In *Artemia* biodiversity studies, RFLP markers have revealed that as a genus*, Artemia* has genetic variability levels that are among the highest within crustaceans [4,21].

More tools for genetic research on *Artemia* are becoming available. Recently, methods for *in vitro* primary cell culture of *A. sinica* have been developed, although cell lines are not available yet [89]. Transgenic *Artemia* have been developed for the production of green fluorescent protein (GFP) and growth hormone (GH) [30]. RNA interference (RNAi) protocols are available for *A. franciscana* [40] and expression studies have been carried out on individual *Artemia* genes as well as on groups of genes during different developmental stages, such as diapause [126,164] and resumption of embryo development after quiescence [53].

Until now, only the mitochondrial genome of several *Artemia* species has been sequenced [208] (Figure 1-9). It has the same coding capacity as most animal mitochondrial genomes and its overall organization is similar to that of *Drosophila* [67].

Most *Daphnia* mitochondrial proteins resemble more those of dipterans (*Drosophila* and *Anopheles*) than those of *Artemia*, both at the nucleotide and the amino acid levels, suggesting that the mitochondrial DNA (mtDNA) of *Artemia* follows an accelerated evolution rate [42].

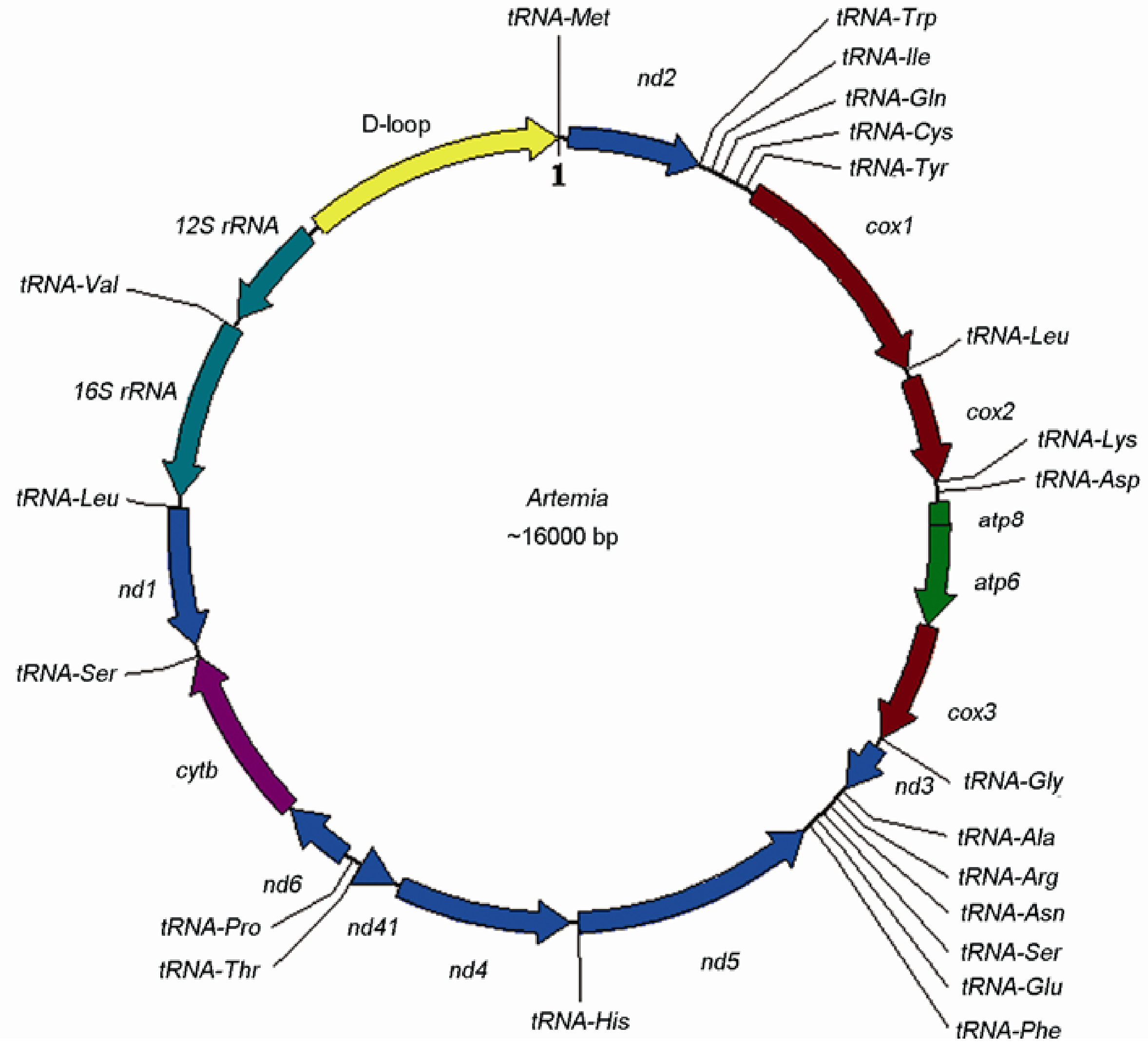


Figure 1-9. Circular physical map of the *Artemia tibetiana* mitochondrial genome [233]. It is ~16 kb long and contains 37 genes, of which 13 protein-coding genes that code for the subunits of the complexes of the mt electron transport system. Genes encoding Complex I (*nd5*, *nd4*, *nd4l*, *nd1*, *nd2*, *nd3*, and *nd6*), Complex III (*cytb*), Complex IV (*cox1*, *cox2*, and *cox3*) and Complex V (*atp8* and *atp6*) are shown in blue, purple, brown, and green, respectively. The other genes are RNA genes, of which 22 are transfer (t)RNA and two are ribosomal (r)RNA (*12S rRNA* and *16S rRNA*) genes (dark green). The major noncoding region (D-loop) is shown in yellow. The clockwise and counter-clockwise arrows indicate genes on the heavy and light strands, respectively.

1.3. Sex determination and sex-determining genes in arthropods

Sex-determining genes activate the developmental program committing the embryo to either the female or the male pathway [13]. Gene regulatory networks that control sex determination vary considerably between and even within animal species and are among the fastest-evolving biological networks [178].

Sex-determining systems have been classified into two main categories: environmental sex determination (ESD) and genetic sex determination (GSD). ESD is initiated by diverse environmental cues and finally triggers the regulation of male or female sex-determining genes. GSD initiates alternate sex-determining developmental pathways by genetic segregation of genes, often located on sex chromosomes. The conventional definition of GSD states that the sex of an individual is fixed upon fertilization by inherited genetic factors. The establishment of the corresponding sexual phenotype is subsequently achieved through sexual development, a process which has classically been divided into sex determination and sex differentiation. Determination is understood as the “master” switch (initial inherited factor) that causes the first steps of the sex determination cascade (usually key sex-determining genes), which activate further downstream genes of sexual differentiation, which in their turn regulate steroid hormone production, eventually leading to one functional gonad type with the corresponding sexual phenotype [79]

Arthropods exhibit a large variety of mostly GSD systems, both at the chromosomal and molecular level. Common GSD systems in arthropods are male heterogamety (XY/XX), female heterogamety (WZ/ZZ) and haplodiploidy, but androdioecy also occurs. In XY/XX systems, females are the homogametic sex, having two sex chromosomes of the same kind[2] (XX), whereas males are the heterogametic sex, having two cytologically distinct sex chromosomes[3] (XY). This system occurs in some insects, such as *Drosophila.* In WZ/ZZ systems, males are the homogametic sex (ZZ) and females are the heterogametic sex (WZ), a system found, for instance, in the insect *Bombyx mori*. Haplodiploidy is a sex determination system in which unfertilized, haploid eggs develop into males and fertilized, diploid eggs develop into females. This system occurs in Hymenoptera (bees, ants, and wasps), Coccidae, Thysanoptera (“thrips”), some spider mites, Homoptera and Coleoptera. An androdioecious sex-determining system is a mix of ZZ males and WZ hermaphrodites and it occurs in crustacean *Eulimnadia texana* [227]

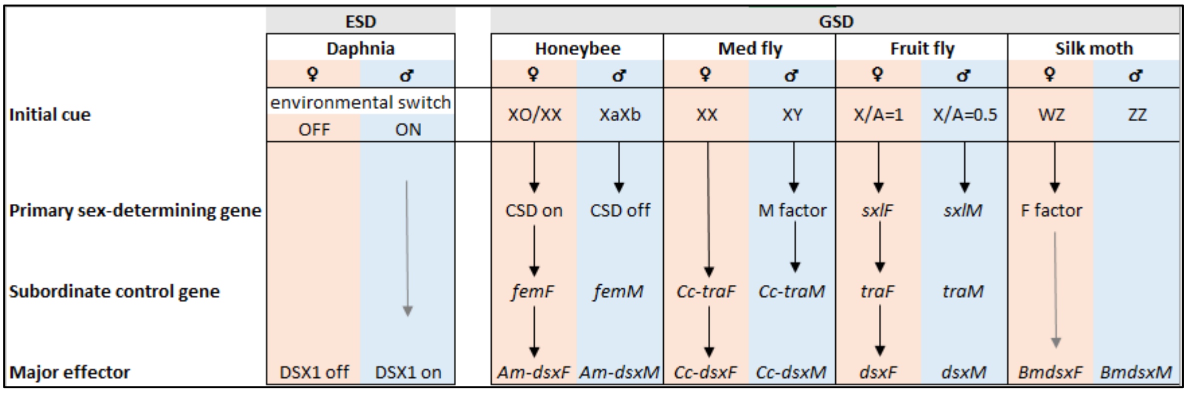
---

[2] Homomorphic sex chromosomes

[3] Heteromorphic sex chromosomes

In most cases of GSD, chromosomal signals activate key sex-determining genes, in their turn triggering the expression of a cascade of downstream sex-determining and further downstream sex-differentiating genes [178]. A simplified sex determination gene cascade is shown for different arthropod sex determination systems (Figure 1-10). In *Drosophila*, the chromosomal signal is provided by the sex chromosome to autosome ratio X/A (X/A=1 signals female development, detailed in Figure 1-11), and by the presence of a Y chromosome in males in other *Diptera and Coleoptera* (such as med fly). In most *Lepidoptera* (WZ/ZZ) [93], sex is signaled by the presence of a W chromosome in females, whereas in *Hymenoptera* (such as *Apis mellifera*), which lack sex chromosomes, sex is signaled by a haplodiploid mechanism [178].

One exception to the GSD rule in arthropods is *Daphnia*, in which the *doublesex* (*dsx*) gene is the sex-determining gene switch, triggered by environmental cues (Figure 1-10). In most other crustaceans, however, sex is determined genetically.



	ESD		GSD							
	Daphnia		Honeybee		Med fly		Fruit fly		Silk moth	
	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂
Initial cue	environmental switch OFF	ON	XO/XX	XaXb	XX	XY	X/A=1	X/A=0.5	WZ	ZZ
Primary sex-determining gene			CSD on	CSD off		M factor	*sxlF*	*sxlM*	F factor	
Subordinate control gene			*femF*	*femM*	*Cc-traF*	*Cc-traM*	*traF*	*traM*		
Major effector	DSX1 off	DSX1 on	*Am-dsxF*	*Am-dsxM*	*Cc-dsxF*	*Cc-dsxM*	*dsxF*	*dsxM*	*BmdsxF*	*BmdsxM*

Figure 1-10. Simplified view of sex-determining pathways in Arthropoda, adapted from Kato *et al*., 2011 [97]. Two different sex determination mechanisms are shown: environmental sex determination (ESD), where the initial cue for sex is an environmental switch (branchiopod crustacean *Daphnia*) and genetic sex determination (GSD), where sex is initially determined by chromosomal signals (insects honey bee (*Apis mellifera*), med fly (*Ceratitis capitata*), fruit fly (*Drosophila*) and silk moth (*Bombyx mori*) [146]. In insects, the initial genetic cues initiate primary sex-determining genes (*CSD*, *Mfactor*, *sxl*, *tra,* or *fem*), which eventually cause the *dsx* gene to be expressed in a sex-specific manner. In *Daphnia*, environmental cues directly switch the major effector *dsx* on or off.

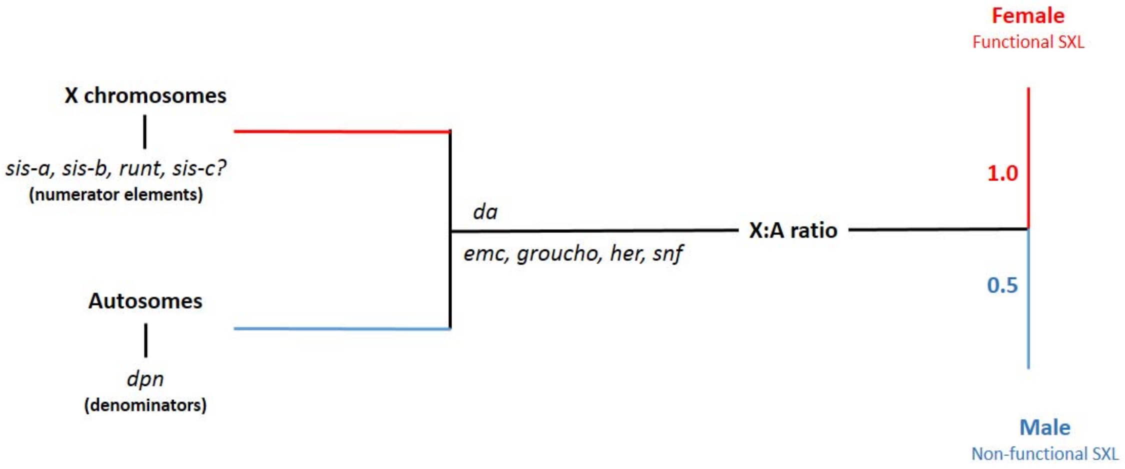


Figure 1-11. The X:A ratio determines sex in *Drosophila*, from Manolakou *et al.,* 2006 [132]. This ratio is communicated through the balance between the X numerator elements (genes *sis-a*, *sis-b*, *runt* and *sis-c*) and the autosomal denominators (*dpn* genes) in conjunction with several maternally derived genes (*emc*, *groucho*, *her* and *snf*). An X:A ratio of 0.5 leads to a non-functional *sex-lethal* protein (SXL) resulting in male development, whereas an X:A ratio of 1 maintains the SXL function, leading to female development.

The sex determination cascade has been elucidated in great detail in the model insect *Drosophila* (Figure 1-12). In *Drosophila* females (X/A=1), the key female-determining gene *Sex-lethal* (*sxl*) is activated by chromosomal signals, causing splicing of one of its immediate downstream target genes *Transformer* (*tra)* to a female-specific RNA-binding protein TRA. This protein controls female-specific splicing of *dsx*, together with *Tra-2* and, ultimately, female-specific transcription factors are produced in specific tissues that control sexual differentiation and behavior [178,202]. In males (X/A=0.5), *sxl* and *tra* are not activated, male-specific *dsx* proteins are produced and male behavior is activated by the male sex-determining gene *fruitless* (*fru*). The *dsx* gene contains two conserved domains: the Dsx/Mab-3 (DM) domain at the N-terminus and the oligomerization domain at the C-terminus [202]. DM domain genes play a related role in *Caenorhabditis elegans* and in vertebrates as well [202].

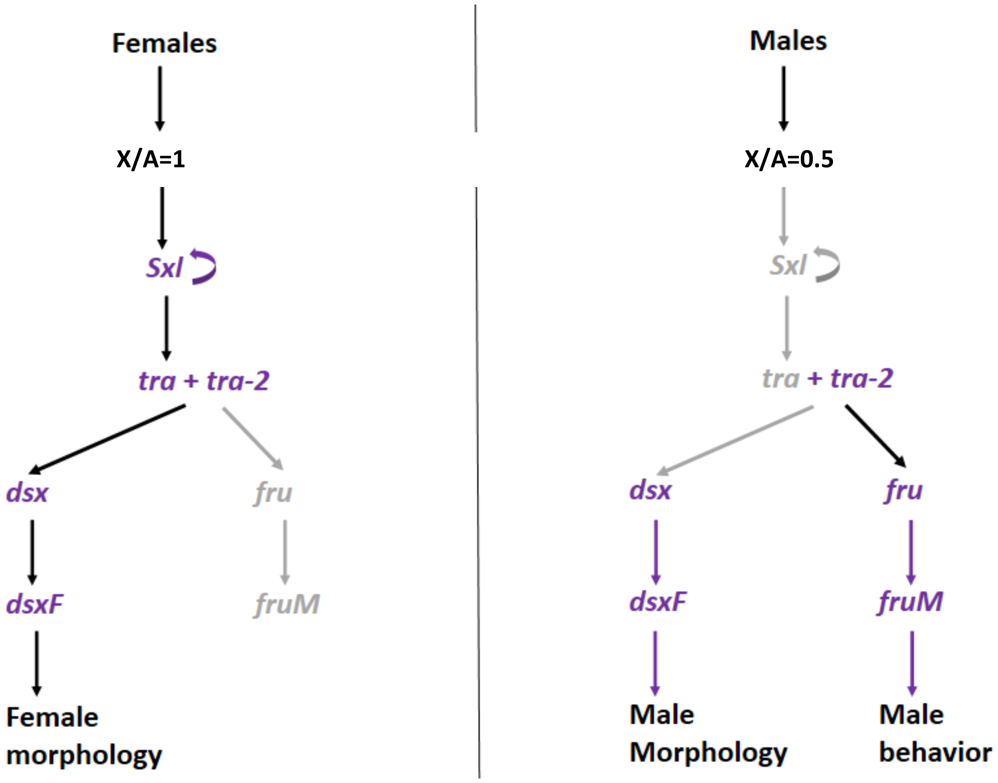


Figure 1-12. Sex determination cascade in *Drosophila*, from Kopp *et al.* 2012 [106]. Inactive gene products and interactions are shown in grey. Primary sex determination occurs early in embryonic development when zygotically transcribed genes located on the X chromosome and the autosomes activate *Sxl* in females (not in males). *Sxl* then sustains itself through a positive autoregulatory loop in females and controls the splicing of *tra*, leading to female-specific splicing of *dsx* transcripts and suppression of male-specific splicing of *fru*. In males, default male-specific splicing of *dsx* and *fru* takes place.

Tra regulation of *dsx* and its taxonomic distribution indicate that the *dsx-tra* axis is conserved in most insects [177,218]. In the female heterogamete *Bombyx mori* (Lepidoptera), *tra* and *tra-2* are probably not involved in sex determination because *Bmdsx (*a homologue of *Drosophila dsx)* has no *Tra/Tra-2* binding motif. A small number of potential candidates for a feminizing gene (*fem*) have been proposed in a limited region of the W chromosome. [62,202].

In conclusion, the primary sex-determining gene at the top of the sex determination cascade (*sxl*, *tra,* or *csd*) has been identified in many insect species, although no sex-determining gene has been identified in insects with female heterogamety, such as *B. mori*. It has been hypothesized that genetic sex determination pathways evolve from bottom to top. As a general rule, the farther upstream in the sex determination cascade, the less structurally conserved the sex-determining genes are (such as *sxl* and *tra*). Despite its conserved function, the *tra* nucleotide sequence, amino acid composition and protein structure vary greatly among insects because they are compatible with different upstream sex-determining signals [218]. Downstream sex-determining genes such as *dsx* (Figure 1-13) and *fru* on the other hand, are highly conserved in a broad range of insects and even in crustaceans [33,63,97].

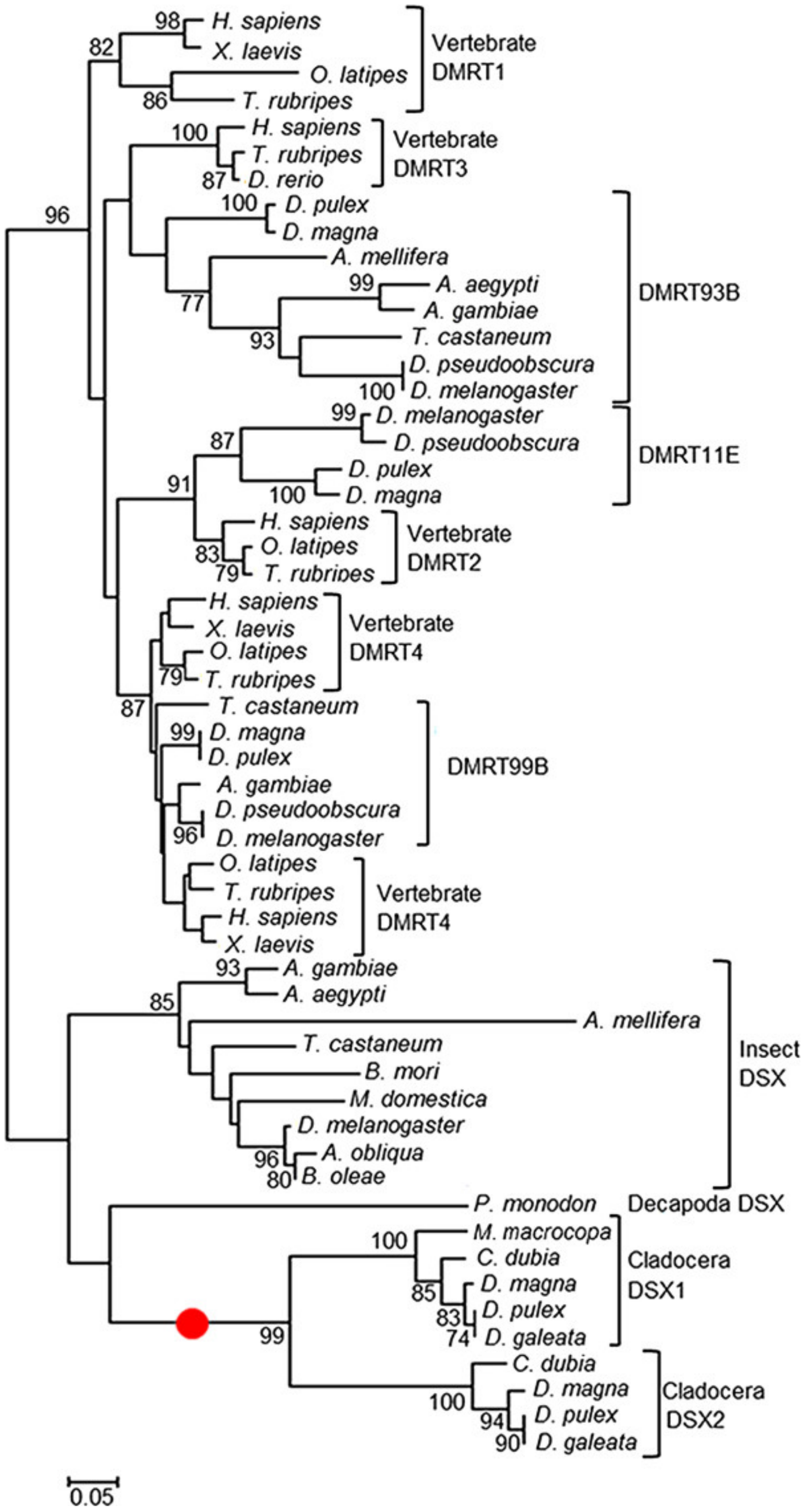


Figure 1-13. Phylogeny of DM domain-containing genes, such as *dsx*, based on amino-acid sequence conservation. DSX in Decapoda (such as *Penaeus monodon*) is similar to that in Cladocera (such as *Daphnia*) [117]. The red dot indicates the duplication period of *dsx* in Cladocera. From [202]

In crustaceans, three major GSD systems have been suggested by cytogenetics and genetic mapping of sex-linked markers: a WZ/ZZ, a XX/XY and an androdioecious sex-determining system (a mix of ZZ males and WZ hermaphrodites, as in *Eulimnadia texana*) [227].

Mainly highly conserved downstream sex-determining genes such as *dsx* have been identified in crustaceans and only potential candidate genes for primary sex determination are known to date [117,124]. More detailed information about sex determination and sex-determining genes in crustaceans is provided in section 3.2.2.2.

In *Artemia*, *s*everal experiments have already hinted at female heterogamety (Chapter 2.2). Large sexual heterochromosomes have been described in *A. salina* and in Yucatán *A. franciscana* [22,193,201]. Female karyotypes of *A. franciscana* show one heteromorphic and 20 homomorphic chromosome pairs (Figure 1-14) [157]. The heteromorphic chromosome pair contains the smallest and the largest chromosome of the complement and has been claimed to correspond to females judging from the chromocentre pattern that these chromosomes exhibit in sexed animals [157].

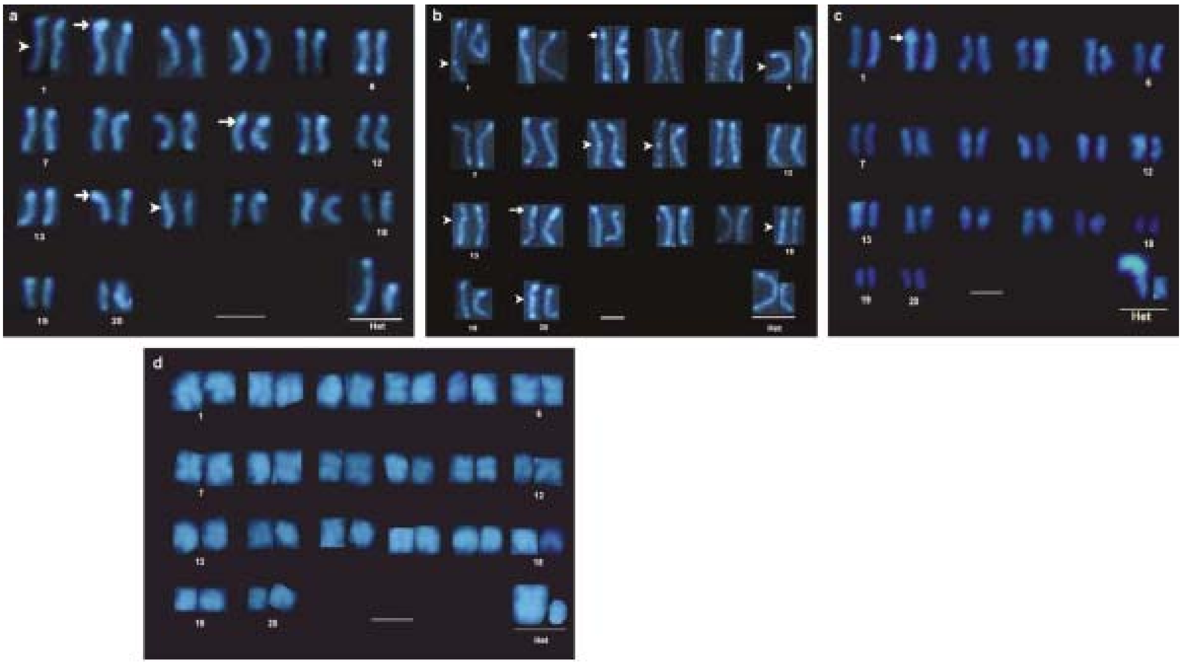


Figure 1-14. Four *Artemia franciscana* karyotypes stained with DAPI [36]: (a) San Francisco Bay-1364, USA, (b) Sosa Texcoco, Mexico, (c) San Francisco Bay-1258, USA and (d) Pichilemu, Chile. Arrows and arrow heads indicate bright fluorescent blocks and less intense heterochromatic blocks, respectively. Bar=5 µm. Heteromorphic chromosomes are underlined. Each *A. franciscana* karyotype (a, b, c and d) shows one heteromorphic and 20 homomorphic chromosome pairs.

In *Artemia*, the sex-determining system and sex-determining genes are currently unknown.

1.4. Research aims

The aims of this PhD thesis were threefold. First, we wanted to construct sex-specific high-density AFLP-based genetic linkage maps for an *Artemia* population, disclosing several sex-linked AFLP markers, assign them to a linkage group and to identify the sex-determining system present in *Artemia*. The produced *Artemia* linkage maps will provide a basis for fine mapping and exploration of the sex-determining region. It will also be a marker resource for mapping genomic loci underlying phenotypic differences among *Artemia* species.

Secondly, with the knowledge of several sex-linked AFLP markers and the sex-determining system in *Artemia*, our goal was to fine map and explore the sex-determining region and to discover one or more putative genes, involved in primary sex determination in *Artemia* with a combined strategy of bulked segregant analysis and next-generation sequencing (NGS). NGS reads from full-sib males and females were used to assemble the *Artemia* genome *de novo*. At this stage, the purpose of the new genome was only to serve as a reference genome for calling "male" and "female" SNPs and as a gene database from which putative sex-determining genes could be selected. The results of this research will help understanding sex determination mechanisms in crustaceans, as well as enable the design of new genetic sexing strategies and, possibly, monosex breeding in different crustacean species of economic importance.

By discovering that the *Artemia* genome assembly met the qualifications to launch a legitimate *Artemia* genome project, we set a new, third aim: to assemble the *Artemia* genome to its best potential and to provide an annotated draft genome assembly representing the core database for current *Artemia* genetic or genomic research. To achieve this goal, most of the known *Artemia* genes should be present in the assembly. The only publicly available annotated crustacean genome is currently that of *Daphnia pulex* [39]. As crustaceans are one of the most diverse arthropod groups (comparable to insects), many, yet unknown, crustacean-specific molecular processes remain to be discovered. The knowledge of the *Artemia* genome would be of high relevance for basic and applied investigation areas, such as stress response, toxicity, crustacean metabolism, biodiversity, *Vibrio* resistance, interaction of sexual and parthenogenetic populations, trophic transfer in a model marine food chain and evolutionary processes in arthropods, such as the evolution of insects.

General scientific and societal impact is to be expected in:

- Aquaculture, because *Artemia* is an important live food and model for other crustaceans used for aquaculture
- Human health, because research on extremophilic features of *Artemia* has revealed genes interesting for cancer research
- Ecosystems, because *Artemia* is a key species of hyper saline ecosystems
- Conservation, because several *Artemia* strains are currently threatened in their natural habitat, among which the San Francisco Bay (SFB) strain of *A. franciscana* studied in this work (personal communication J. Clegg).

The following chapters describe the production of *Artemia* AFLP-based genetic maps with several sex-linked AFLP markers and the identification of the sex-determining system in *Artemia* (Chapter 2), the discovery of putative *Artemia* sex determination genes with bulked segregant analysis (Chapter 3) using the *de novo* assembly and annotation of the *Artemia* genome (Chapter 4). In the final chapter (Chapter 5), research results are discussed and potential future research is proposed.

Chapter 2

A first AFLP-based genetic linkage map for brine shrimp *Artemia franciscana* and its application in mapping the sex locus

---

Chapter 2 is based on the paper:

De Vos S, Bossier P, Van Stappen G, Vercauteren I, Sorgeloos P, Vuylsteke M (2013) A first AFLP-based genetic linkage map for brine shrimp *Artemia franciscana* and its application in mapping the sex locus. PLoS One, 8 (3): e57585.

2.1. Abstract

We report on the construction of sex-specific linkage maps, the identification of sex-linked markers and the genome size estimation for the brine shrimp *Artemia franciscana*. Overall, from the analysis of 433 AFLP markers segregating in a 112 full-sib family, we identified 21 male and 22 female linkage groups (2*n* = 42), covering 1,041 and 1,313 cM respectively. Fifteen putatively homologous linkage groups, including the sex linkage groups, were identified between the female and male linkage map. Eight sex-linked AFLP marker alleles were inherited from the female parent, supporting the hypothesis of a WZ/ZZ sex-determining system. The haploid *Artemia* genome size was estimated to 0.93 Gb by flow cytometry. The produced *Artemia* linkage maps provide the basis for further fine mapping and exploring of the sex-determining region and are a possible marker resource for mapping genomic loci underlying phenotypic differences among *Artemia* species.

2.2. Introduction

Artemia, known as brine shrimp, is a genus of small planktonic crustaceans found worldwide in natural salt lakes and salterns [109]. Their larvae (nauplii) are the most commonly used live food in aquaculture activities, specifically for larval growth of more than 85% of the marine species reared in aquaculture [98,111]. Adult *Artemia* survive extreme salinities, while their encysted embryos (cysts), produced under stressful conditions, have a unique tolerance for high doses of UV and ionizing radiation, anoxia, thermal extremes and desiccation-hydration cycles [12,23,172]. Cysts remain viable for years and produce nauplii within 24 h after hydration.

An overview of *Artemia* cytogenetics, DNA content and available molecular tools is provided. Six different sexually dimorphic species can be found in the *Artemia* genus, among which *Artemia franciscana* Kellogg, 1906 [12] and several obligate parthenogenetic *Artemia* populations ranging in ploidy from 2n to 5n [131]. All sexually dimorphic *Artemia* species are diploids with 2n = 42, except *A. persimilis* (2n = 44) [157]. The *Artemia* genome size has been assessed with two different techniques producing discordant estimates: 2.93 Gb (3 pg) by Feulgen densitometry [171] and 1.47 Gb (1.5 pg) by DNA reassociation kinetics [212]. Despite the use of flow cytometry in the most recent evaluations of crustacean genome sizes [121,167,168], so far no flow cytometry-based estimates of the *Artemia* genome have been published. To date, genomic resources for *Artemia* have been limited to RAPD [14,25], RFLP [21], AFLP [197,204], microsatellite markers [144] and the 15,822 bp mitochondrial genome sequence [196,208].

In crustaceans, three major genetic sex determination systems have been suggested by cytogenetics and sex-linked markers: WZ/ZZ (females are the heterogametic sex), XX/XY (males are the heterogametic sex) and androdioecy (a mix of ZZ males and WZ hermaphrodites, as in *Eulimnadia texana*) [227]. Examples of crustaceans with an XX/XY sex-determining system are decapods such as the Chinese mitten crab *Eriocheir sinensis* [59,155], terrestrial isopods and the amphipods *Orchestia cavimana* and *O. gammarellus* [91]. However, a WZ/ZZ sex-determining system has been found in decapods such as *Litopenaeus vannamei* [8], tiger shrimp *Penaeus monodon* [192], *Macrobrachium rosenbergii* [215], kuruma prawn *Penaeus japonicus* [119], Australian red claw crayfish *Cherax quadricarinatus* [155] and in isopods like *Armadillidium vulgare* and all Valvifera, except *Saduria entomon* [200].

In bisexual *Artemia*, female heterogamety has been suggested previously by observation of sexual heterochromosomes in *A. salina* [193], *A. franciscana* and *A. persimilis* [157]; by crossing experiments with *A. franciscana* showing a recessive sex-linked trait called “white eye” [22] and by karyotyping and heterochromatin experiments showing one heterochromatic block in female and two in male *A. persimilis* [157].

Over the last decade, linkage maps have been developed for a number of crustaceans such as *Daphnia pulex* [43]*, D. magna* [175], *Tigriopus californicus* [58], *P. monodon* [130,228,230], *L. vannamei* [52,158,234], *Fenneropenaeus chinensis* [120,226] and *P. japonicus* [118]. Sex-linked markers have been found in males of the isopod *Mysis relicta* [207] and of *Triops cancriformis* [133]. In female crustaceans, sex-linked markers have been found in the isopods *Paracerceis sculpta* [186] and *Jaera ischiosetosa* [187], in the crab *Cancer setosus* [71], in penaeid shrimps *L. vannamei* [234] and *P. monodon* [192] and in giant freshwater prawn *M. rosenbergii* [214]. Moreover, a hermaphrodite-determining allele has been studied in the androdioecious branchiopod *E. texana* [153]. So far, neither linkage maps, nor trait-linked markers including sex-linked markers have been identified in *Artemia* [157].

Genetic linkage maps are invaluable in forward genetic analyses for the identification of the genomic loci responsible for phenotypic differences. From this perspective, *Artemia* offers a number of major advantages for time-effective generation of experimental mapping populations and for mapping of natural allelic variation. They have a short generation time (2-4 weeks), offspring production of several hundred individuals per brood, storability of cysts for years, easy breeding in large numbers and levels of genetic variability that are among the highest within crustaceans [4,21,234]. In addition, we expect that forward genetic approaches in *Artemia* are not only restricted to *Artemia*-specific traits, but are also valuable for mapping traits such as sex, *Vibrio* pathogen resistance and growth rate, segregating in commercially important crustaceans. We believe therefore, that *Artemia* could be a useful model species for other crustaceans.

In the present study, we report on a first AFLP-based linkage map of *A. franciscana*. We additionally present eight sex-linked markers that disclose the linkage group corresponding to the W chromosome and confirm female heterogamety in *A. franciscana*. Finally, we report on the estimation of the *A. franciscana* genome size by flow cytometry.

2.3. Materials and methods

Genetic mapping techniques are explained in section 1.2.1.1.

2.3.1. Mapping population

Cyst material of the *A. franciscana* strains from San Francisco Bay, USA (SFB; ARC1364) and Vinh Chau, Vietnam (VC; ARC1349) was obtained from the Laboratory of Aquaculture & *Artemia* Reference Center cyst bank (http://www.aquaculture.ugent.be). The SFB strain was first introduced in Vinh Chau, Vietnam in 1982, eventually resulting in the new VC strain in the late 1980`s [95]. First, cysts from both strains were hatched separately in aerated seawater (28 °C, salinity 35 g l^{-1}). The instar I nauplii of each strain were then harvested and reared for a week in aerated seawater with added sea salt (Instant Ocean®, 28 °C, final salinity 70 g.l^{-1}) and fed with *Tetraselmis suecica*, a marine unicellular green alga. The *Artemia* were subsequently transferred to individual Falcon tubes and kept there under the same conditions for seven days until sexual maturation. One controlled cross between VC (♀) and SFB (♂) was then made, resulting in full-sib F_1 progeny that was collected over a sieve every two days and grown until maturity under the same conditions as the parental generation. Adult full-sib progeny was rinsed with sterile distilled water and the phenotypic sex of each offspring individual was determined visually (Figure 1-3.). For gut evacuation before DNA extraction, the offspring and parents were starved during 24 h, followed by removal of the brood pouch in females. Parents and progeny were stored individually at -20 °C.

2.3.2. DNA extraction

DNA was extracted from parents and their 112 F_1 offspring according to a modified CTAB-method for shrimp tissue [80]. Briefly: to each sample, ground in liquid N_2, 150 µl of CTAB buffer was added. After homogenization, 750 µl of extra CTAB buffer was added and the mix was left at 25 °C for 30 min. PCA solution was added (600 µl; 25:24:1 phenol/chloroform/isoamylalcohol). After centrifugation, 800 µl of the upper aqueous phase was added to 600 µl of CA solution (24:1 chloroform/isoamylalcohol) and the mix was homogenized. To 700 µl of the upper aqueous phase, 630 µl of isopropanol was added. The mix was incubated for 1 h at -70 °C. After centrifugation, the pellet was washed with 600 µl of ethanol 70%, air-dried in a 60 °C oven and resuspended in 20 µl of sterile distilled water. DNA quality and concentration were assessed on a 1% agarose gel.

2.3.3. Segregation analysis and linkage mapping

AFLP analysis with fluorescent dye detection was performed on a LI-COR long read-IR^2 4200 (LI-

COR Biosciences) as described by Vuylsteke *et al*. [225]. Briefly, AFLP template fragments were generated by 1) digestion of genomic DNA with a combination of the two restriction enzymes *Eco*RI and *Mse*I, followed by 2) ligation of the double-stranded *Eco*RI- and *Mse*I- specific adapters to the fragment ends. A pre-amplification step was performed to PCR amplify subsets of the EcoRI/*Mse*I templates, using primers that match the adapter sequences, carrying one selective nucleotide at their 3'-end. Finally, 65 selective PCR-amplifications were performed using three selective nucleotides added to the *Eco*RI and *Mse*I primers. The *Eco*RI+3/*Mse*I+3 primer combinations (PCs) are listed in Table 2-1. AFLP analysis of parents and 112 offspring was done on two separate 64-lane gels per PC.

Table 2-1. List of the 65 primer combinations used for AFLP analysis. E: *Eco*RI primer with three selective bases; M: *Mse*I primer with three selective bases (1, 2, 3, 4 correspond to A, C, G, T).

E112M112	E112M222	E112M323	E112M431
E112M113	E112M223	E112M331	E112M432
E112M121	E112M224	E112M332	E112M433
E112M122	E112M231	E112M333	E113M112
E112M123	E112M232	E112M334	E113M114
E112M124	E112M233	E112M341	E113M122
E112M131	E112M234	E112M342	E113M123
E112M132	E112M241	E112M343	E113M132
E112M133	E112M242	E112M344	E113M142
E112M134	E112M243	E112M411	E113M143
E112M142	E112M244	E112M412	E113M211
E112M143	E112M311	E112M413	E113M212
E112M211	E112M312	E112M414	E113M213
E112M212	E112M313	E112M421	E113M214
E112M213	E112M314	E112M422	
E112M214	E112M321	E112M423	
E112M221	E112M322	E112M424	

The degree of polymorphism between the two parental strains was estimated based on AFLP fragments amplified by four PCs (E112M212, E112M213, E112M233 and E112M234). These estimations were used to estimate nucleotide diversity for each PC according to methods of Innan *et al*. [84].

AFLP markers were scored using the specific image analysis software AFLP-Quantar*Pro* (http://www.keygene-products.com) as described in Vuylsteke *et al*. [225].

Each AFLP marker was identified by (1) a code referring to the corresponding PC (Table 2-1), followed by (2) the relative molecular size of the fragment in nucleotides as estimated by AFLP-Quantar*Pro* by comparison with the used DNA marker SmartLadder SF 100-1000 bp, and (3) a tag referring to the type of marker. Parental AFLP markers segregating 1:1 in the F_1 progeny were heterozygous in either the female (female marker, tagged as “F”) or the male parent (male marker, tagged as “M”) and homozygous absent in the other parent. AFLP markers heterozygous in one of the parents and homozygous present in the other were not included in the linkage analysis, because heterozygotes could not be reliably discriminated from individuals homozygous for the “band present” allele. No tag was used for biparental markers, which are heterozygous in both parents and thus, segregate 1:2:1 in the F_1 progeny. Parental and biparental AFLP markers were scored co-dominantly based on relative fragment intensities resulting in more genetic information compared to dominant (present/absent) scoring and hence, speeding up the mapping process [225]. Biparental markers were, however, scored dominantly when the heterozygotes could not reliably be discriminated from the individuals homozygous for the “band present” allele.

Linkage and segregation analyses were performed using the software package Joinmap 4 [211]. The mapping population type was set to CP (i.e. a population resulting from a cross between two heterogeneously heterozygous and homozygous diploid parents, linkage phases originally unknown). The segregation type was encoded according to Joinmap 4 recommendations [211]. A logarithm of the odds (LOD) threshold range between 2.0 and 14.0 was initially used to group parental markers. Only linkage groups containing at least three markers were considered for map construction. Segregation distortion of markers was tested by using a χ^2-test as implemented in Joinmap 4. Graphical presentation of linkage groups was done with the software MapChart [221].

2.3.4. *Artemia* genome size estimation by flow cytometry

The haploid genome size (GS) of *Artemia* was assessed against the rainbow trout (haploid GS 2.4-3.0 pg or 2.35-2.93 Gb [69]) and the chicken genome (haploid GS 1.07 pg or 1.05 Gb [139]), both used as internal standards.

The consistency of the used method was assessed by calibrating rainbow trout nuclei (2 µl of freshly drawn heparinized *Oncorynchus mykiss* blood) against chicken erythrocyte nuclei (2 µl of 10x diluted BioSure®CEN singlet, *Gallus gallus domesticus*, Rhode Island Red female).

Each of the four *Artemia* individuals (i.e. four full-sib males from the VC (♀) x SFB (♂) cross) were chopped together with internal standard material using a razor in 1 ml of Galbraith`s buffer as described in Dolezel and Bartos [51]. Cell suspensions were filtered through a 30 µm mesh, put on ice and nuclei were co-stained in the dark for 2 min with 50 µl of fluorescent DNA stain

Propidium Iodide (Sigma-Aldrich PI solution in water 1 mg/ml). The use of PI staining on *A. franciscana* (GC% 32) [44], *O. mykiss* (GC% 42) [69] and *G. domesticus* (GC% 47) [239] was chosen to avoid a GC content-linked bias, as occurs with DAPI staining [51]. At least 5,000 nuclei were analyzed for each co-stained sample, using a Modular Flow cytometer and cell sorter (MoFlo Legacy, Cytomation) with a 488 nm Argon laser and PI emission bandpass filter of 580/30 nm. Instrument calibration was performed using Flow-check Fluorospheres (Beckman Coulter) and internal standards. Fluorescence of the nuclei was recorded linearly with the software Summit v4.3. For each co-stained sample, fluorescence histograms were generated and mean fluorescence values were calculated with the flow cytometry data analysis software Cyflogic 1.2.1. The haploid GS of for each *Artemia* sample was calculated according to the following formula [167]:

$$GS = \frac{F_s \times F_{is}}{F_{is}}$$

where F_s is the mean fluorescence of the sample and F_{is} is the mean fluorescence of the internal standard.

2.4. Results

2.4.1. Segregation analysis and linkage mapping

A total of 65 AFLP PCs resulted in a total of 531 markers, of which 433 were parental (239 female, 194 male) and 98 markers were biparental. Based on only four primer combinations (PCs) yielding 180 AFLP fragments, 36% of the fragments segregated between both parents. The average *A. franciscana* nucleotide diversity was 6.2%.

First, a parental map including only parental markers was constructed. Summary statistics for the parental maps are listed in Table 2-2. The grouping of parental markers at a LOD score ranging from 5.0 to 6.0 resulted in a number of linkage groups corresponding with the haploid chromosome number (n=21). The female map, containing 225 markers (Figure 2-1), resulted in 22 “female” linkage groups (LG) spanning 1,312.9 cM; the male map, containing 181 markers (Figure 2-2), resulted in 21 “male” LG spanning 1,041.3 cM.

Twenty-eight percent of the analyzed parental markers showed significant ($p < 0.05$; χ^2 test) segregation distortion. Male markers were more often distorted than female markers (31% resp. 25%). Some larger genomic regions did not contain any markers (e.g. 32.5 cM in LG Female_6, Figure 2-1; 38.0 cM in LG Male_2, Figure 2-2); despite the low median inter-marker distances of 3.9 and 3.1 cM for the female and the male linkage map (Table 2-2).

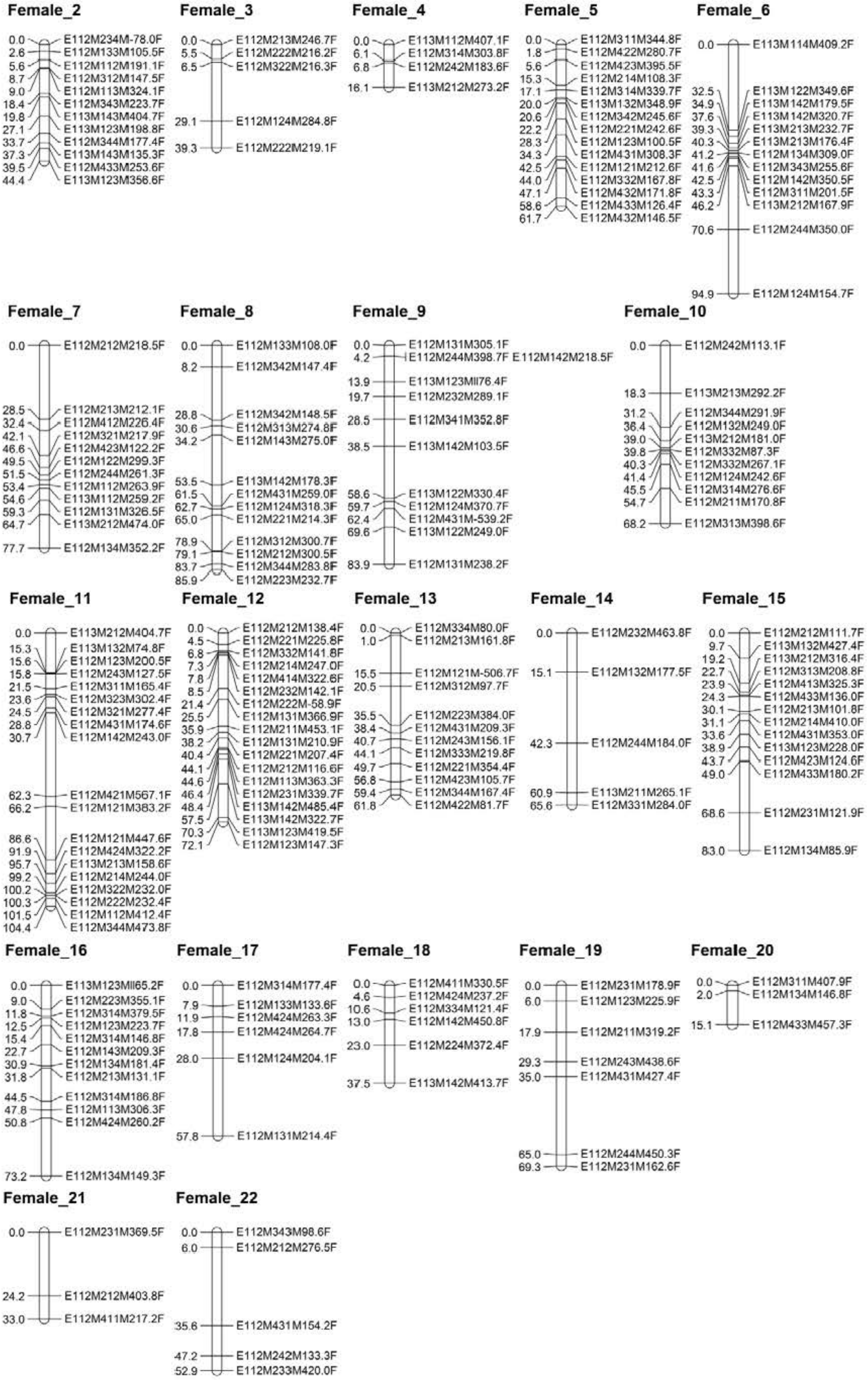


Figure 2-1. ***A. franciscana*** **autosomal female linkage groups. Twenty-one linkage groups representing the** ***A. franciscana*** **autosomal genome containing markers originating from female parental strain Vinh Chau (ARC1349). Each AFLP marker is represented by (1) a code referring to the corresponding PC (Table 2-1), (2) the molecular size of the fragment in nucleotides ("M") and (3) the type of parental marker (female marker, tagged as "F"). Cumulative marker distances (cM) are indicated on the left.**

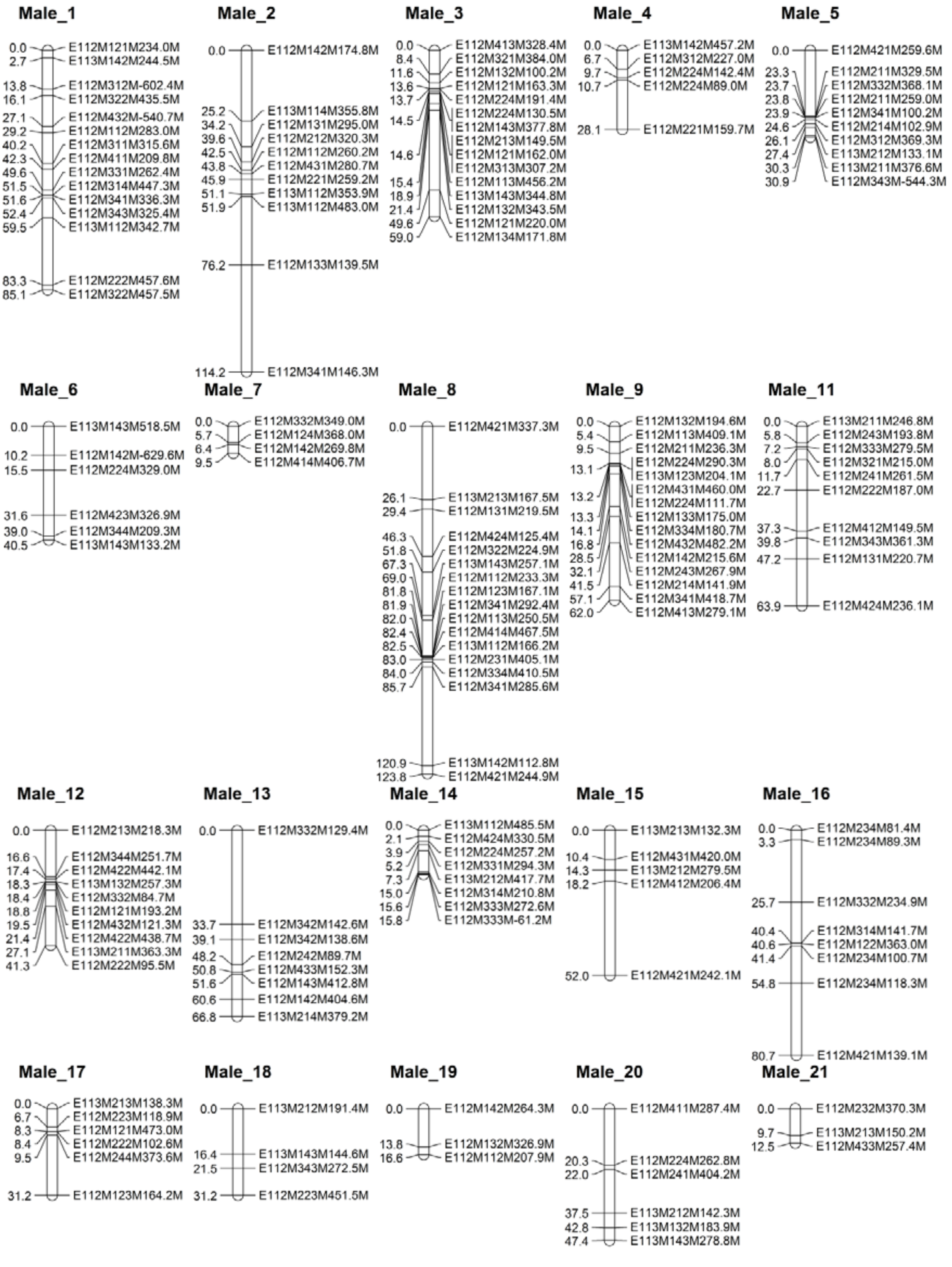


Figure 2-2. ***A. franciscana*** **autosomal male linkage groups. Twenty linkage groups representing the** ***A. franciscana*** **autosomal genome containing markers originating from male parental strain San Francisco Bay (ARC1364). Each AFLP marker is represented by (1) a code referring to the corresponding PC (Table 2-1), followed by (2) the molecular size of the fragment in nucleotides ("M") and (3) the type of parental marker (male marker, tagged as "M"). Cumulative marker distances are indicated on the left (cM).**

Table 2-2. Statistics for the female and male linkage maps.

		Female (Vinh Chau)	Male (San Francisco Bay)
No. of linkage groups		22	21
No. of markers mapped per linkage group	Min	3	3
	Max	19	17
	Median	12	8
	Mean	10	9
	Total	225	181
Size of linkage groups (cM)	Min	15.1	9.5
	Max	104.4	123.8
	Median	63.7	44.4
	Mean	59.7	52.1
	Total	1312.9	1041.3
Intermarker distance (cM)	Min	0.0	0.0
	Max	32.5	38.0
	Median	3.9	3.1
	Mean	6.5	6.6

Next, an integrated map was created including the 98 biparental markers and 406 previously mapped parental markers (Figure 2-3). By including biparental markers, groups consistent with linkage groups of the parental map were obtained at a LOD threshold ranging between 6 and 10. Sixty-nine percent of the biparental marker loci showed significant segregation distortion ($p < 0.05$; χ^2 test). These loci were still included in map construction and evaluated for quality afterwards, since significant segregation distortion is inherent to relatively small experimental mapping population sizes of ~100 individuals. Forty-nine biparental markers (50%) could be mapped in the female as well as in the male map, identifying 15 homologous linkage groups including the sex linkage groups (Figure 2-3 and Figure 2-4).

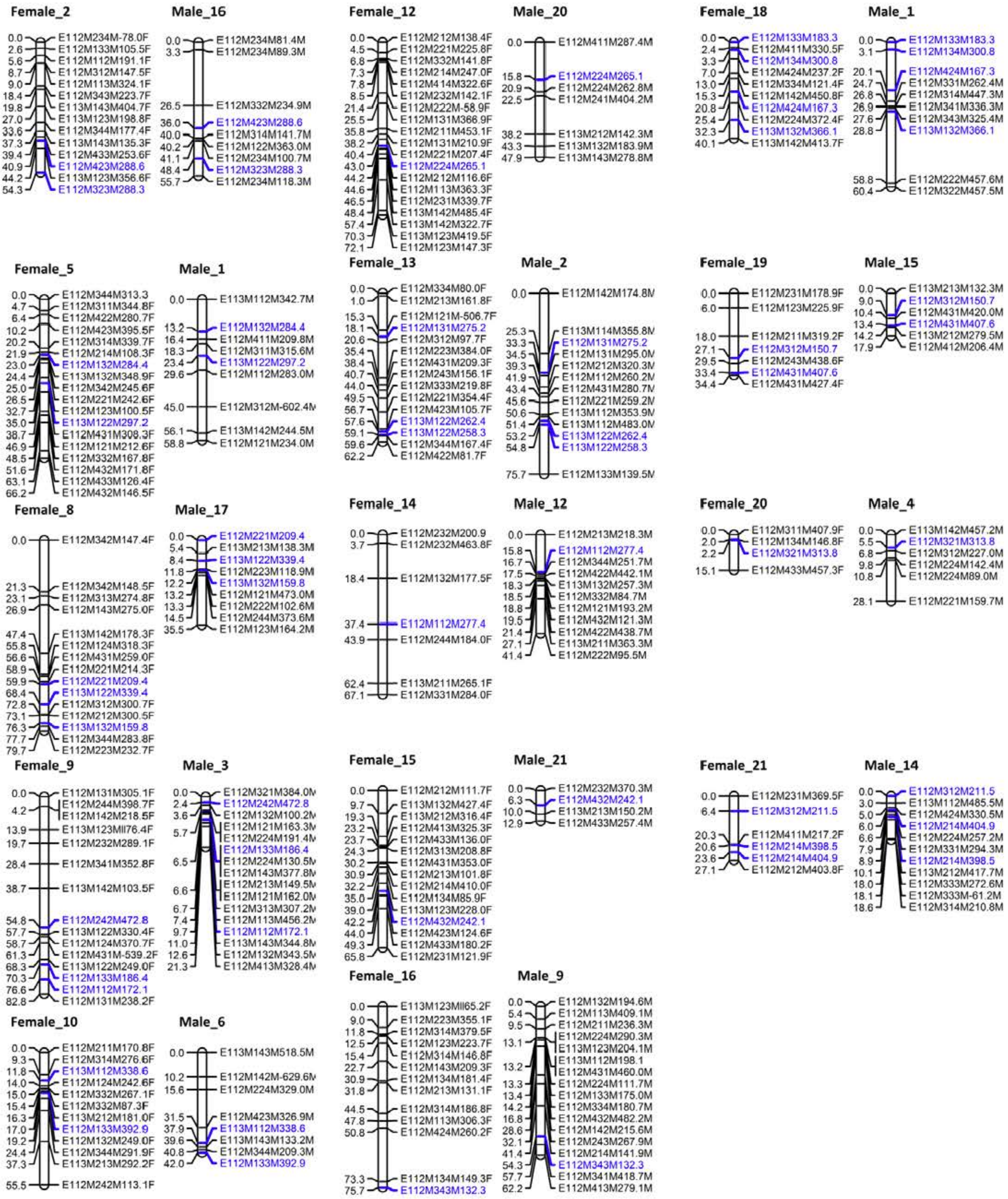


Figure 2-3. *Artemia franciscana* homologous autosomal female and male linkage groups. Fourteen homologous autosomal linkage group pairs. Each AFLP marker was identified by (1) a code referring to the corresponding PC (Table 2-1), followed by (2) the molecular size of the fragment in nucleotides ("M") and (3) the type of marker (female marker, tagged as "F", male marker, tagged as "M", biparental marker, no tag). Common biparental markers are indicated in blue. Cumulative marker distances are indicated on the left (cM).

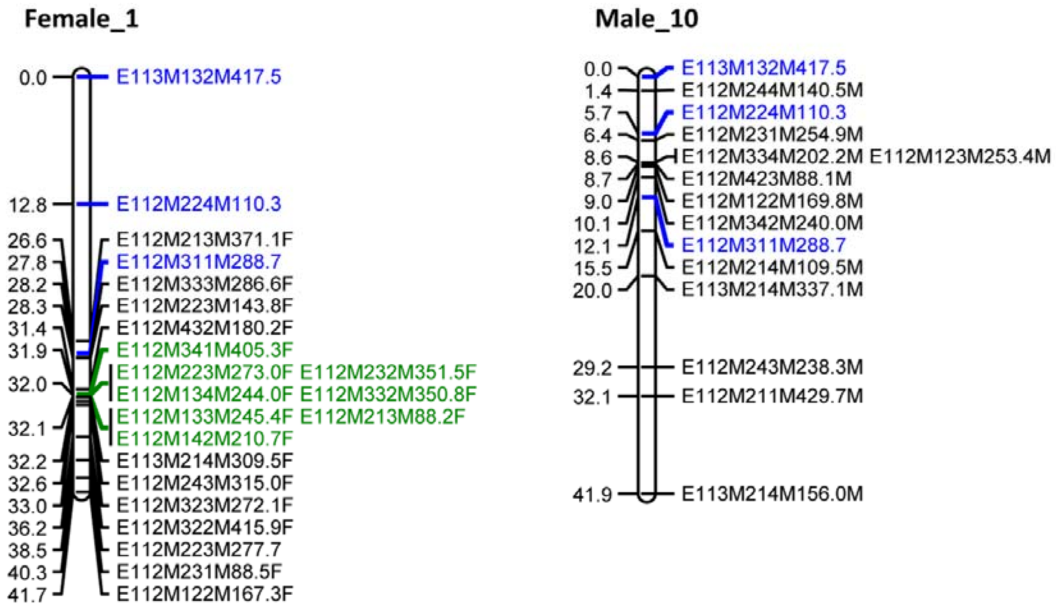


Figure 2-4. ***Artemia franciscana*** **sex linkage groups. Female linkage group Female_1 corresponds with the W chromosome. The homologous male linkage group Male_10 corresponds with the Z chromosome. Each AFLP marker is represented by (1) a code referring to the corresponding PC (Table 2-1), followed by (2) the molecular size of the fragment in nucleotides and (3) the type of marker (female marker, tagged as "F", male marker, tagged as "M", biparental marker, no tag). Common biparental markers are indicated in blue. Markers fully linked to sex are marked in green. Cumulative marker distances (cM) are indicated on the left.**

2.4.2. Linkage mapping of the sex locus

Staelens *et al.* [192] described segregation patterns of sex-linked AFLP markers that unequivocally differentiate the WZ/ZZ and XX/XY sex-determination system (Table 2-3).

Table 2-3. Segregation patterns of sex-linked AFLP markers. WZ–ZZ, patterns 1–5; XX–XY, patterns 6–10. The A allele represents the "AFLP band present" allele. Offspring genotype classes as observed for segregation patterns 4 and 9 cannot unambiguously discriminate between the two sex determination mechanisms. Offspring genotype classes observed for patterns 5 and 10 are identical to those expected for non-sex-linked markers. Hence, observations of patterns 4, 5, 9 and 10 are not informative in characterizing the sex-determination system. Segregation patterns unique for the two sex-determination mechanisms are indicated in blue. Adapted from Staelens *et al.* (2008) [192].

Pattern	Dam genotype	Sire genotype	Genotype of female offspring	Genotype of male offspring
WZ–ZZ				
1	Aa	aa	Aa	aa
2	aA	aa	aa	Aa
3	aA	AA	aA	AA
4	aA	aA	Aa or aa	AA or Aa
5	aa	Aa	Aa or aa	Aa or aa
XX–XY				
6	aa	aA	aa	Aa
7	aa	Aa	Aa	aa
8	AA	Aa	AA	Aa
9	Aa	Aa	AA or Aa	Aa or aa
10	Aa	aa	Aa or aa	Aa or aa

We observed eight AFLP markers, spanning a region of 0.2 cM on LG Female_1 (markers in green, Figure 2-4) segregating according to pattern 1 and a single marker (E112M122M167.3F) according to pattern 2. Both segregation patterns are expected under the assumption of female heterogamety. None of the 433 AFLP markers segregated according to patterns 6, 7 and 8, expected under the assumption of male heterogamety. This could be concluded from mapping of the sex phenotype by including it as a single marker segregating as a female or male- specific marker, indicated as SEX1 and SEX2, respectively. SEX1 could be assigned to the linkage group Female_1. In contrast, SEX2 did not cosegregate with any of the markers of the linkage groups.

The male linkage group Male_10 was identified as homologous to Female_1 (Figure 2-4). In conclusion, the mapping of the sex phenotype and the observed segregation patterns of actual sex-linked AFLP markers strongly favour female over male heterogamety in *Artemia*.

2.4.3. *Artemia* genome size estimation by flow cytometry

Using trout blood as the internal standard, the haploid female chicken genome size (GS) determined by flow cytometry was 1.05 Gb (1.07 pg) as previously reported for female chicken [139]. We preferred rainbow trout nuclei as the internal standard in the assessment of the *Artemia* GS because their fluorescence values did not overlap with those of *Artemia*, as was the case with fluorescence values obtained from chicken nuclei. Using rainbow trout nuclei as the internal standard, the *A. franciscana* haploid genome size was estimated to 0.93 ± 0.09 Gb (0.97 ± 0.09 pg; n=4). Fluorescence histograms for each sample and for chicken are shown in Figure 2-5. Fluorescence peaks were relatively broad due to cell debris from the previously frozen *Artemia* individuals, but average DNA content estimates were consistent throughout the different samples, shown by the small standard error.

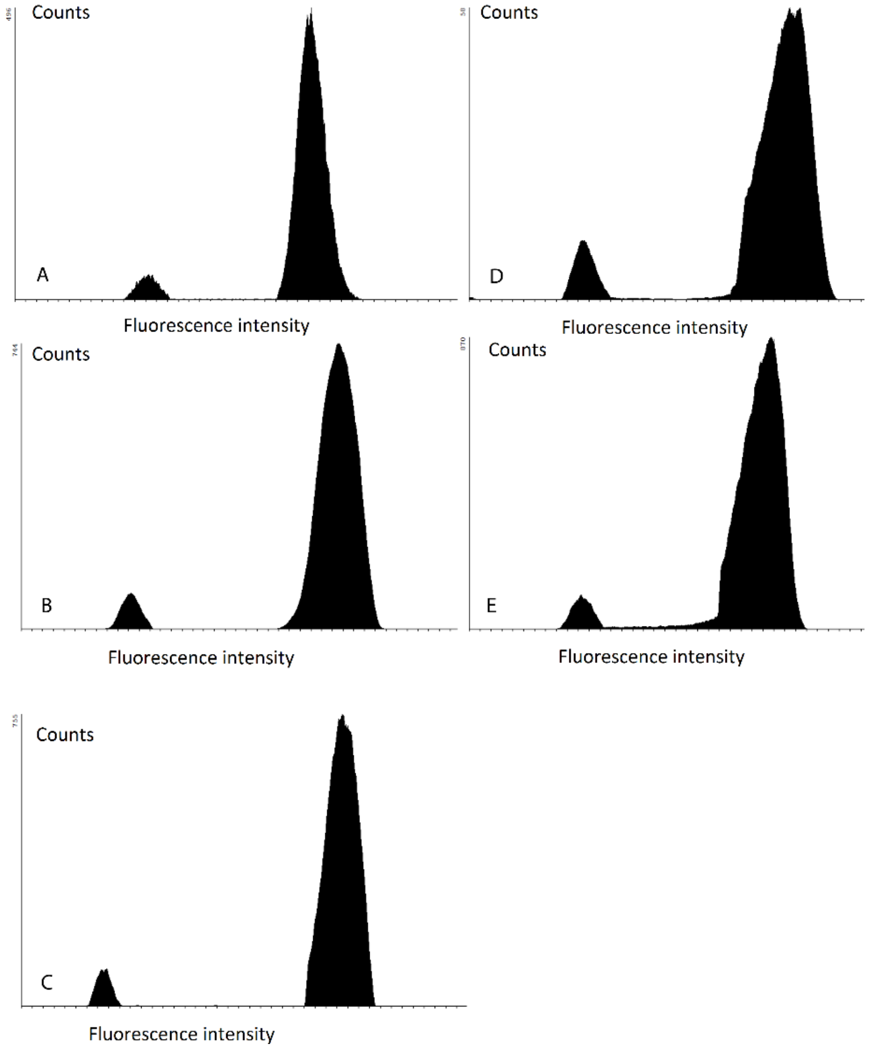


Figure 2-5. Fluorescence histograms for *Artemia franciscana* and chicken DNA content estimation. Nucleus counts for different fluorescence intensities (linear scale) for four different *A. franciscana* male individuals with trout as the internal standard (A, B, C and D) and for chicken CEN with trout as the internal standard (E).

2.5. Discussion

We present the first sex-specific AFLP linkage maps and sex-linked markers as well as a consistent genome size (GS) estimation for the brine shrimp *A. franciscana*.

The linkage analysis of 433 parental AFLP markers segregating in a 112 full-sib family identified 21 male and 22 female linkage groups, corresponding very well with the haploid chromosome number in *A. franciscana* (2n = 42) [157]. Most likely, the markers in small linkage groups (LG) such as Female_20 (Figure 2-1) would join one of the other 21 LG by adding more markers to the female map. More female than male markers were generated, suggesting that maternal *A. franciscana* strain Vinh Chau (VC) has more unique alleles compared to paternal strain San Francisco Bay (SFB). This seems a logical consequence of the SFB origins of VC. The level of polymorphism between the two *A. franciscana* parental strains was estimated at 36%, which is in the range of 9-50% estimated previously by Kappas *et al*. [95].

Given their high marker density, the produced genetic maps are adequate for the anchoring of *Artemia* genome sequences to facilitate the future construction of physical maps for each of the 21 chromosomes. This will be especially useful, considering the numerous reports of repetitive sequences in *Artemia* [26]. *Artemia* linkage maps will also allow future linkage studies in *Artemia* for important crustacean traits such as resistance to *Vibrio*, the most common bacterial pathogen in worldwide marine fish and shellfish aquaculture.

Fifteen homologous linkage groups, including the LG representing the sex chromosomes, were identified between the female and male linkage maps by including biparental markers in the linkage analysis. This study identified eight sex-linked AFLP marker alleles mapping to one locus and inherited from the female parent, suggesting *A. franciscana* adopts a genetic WZ/ZZ sex-determining system. *Artemia* sex-linked markers will enable the study of nauplii sex ratios and their dynamics in natural *Artemia* populations. They will also enable the further fine-mapping of the sex-determining locus and the subsequent identification of the primary sex-determining gene(s). Furthermore, based on sequence homology with *Artemia*, sex-determining genes might be identified in commercially valuable crustaceans, enabling PCR-based allele-specific assay development in the framework of mono-sex culture development in shrimp [216].

The clustering of eight sex-linked markers in a 0.2 cM region suggests reduced recombination, which is often found in sex-linked regions [18]. Genes from a region that stopped recombining in the early evolution of sex chromosomes have a high sequence divergence, allowing an estimate of when the W and Z chromosomes first stopped recombining and thus, the age of the sex chromosome system [18].

The estimated *Artemia* GS in this study (0.93 Gb) is smaller than earlier estimates: 2.93 Gb by Feulgen densitometry [171] and 1.47 Gb by DNA reassociation kinetics [212]. "*A. salina"* used to be a general name for all *Artemia* species, presently confounding the identity of the investigated *Artemia* in many studies [109]. Because the *Artemia* DNA content measured by Feulgen densitometry on "*A. salina*" is almost a twofold of that measured by DNA reassociation kinetics, Feulgen densitometry might have been performed on a tetraploid *A. parthenogenetica*, as suggested by Vaughn [212]. Also, the absolute *A. franciscana* karyotype size varies between 60.68 µm and 139.26 µm [157], showing that significant intra-specific variation in DNA content could explain the high Feulgen densitometry values as well.

Vaughn [212] calculated the *Artemia* haploid GS by DNA reassociation kinetics, based on an *A. franciscana* GC content of 42%. More recent measurements however, show an *A. franciscana* (SFB) GC content of 32% determined by CsCl centrifugation and confirmed by direct chemical analysis and renewed thermal denaturation [44]. An estimated GC content lowered by 1% results in a 0.018% lower haploid DNA content estimated by DNA reassociation kinetics [183]. Hence, based on a GC content of 32%, the corrected *A. franciscana* DNA content estimated by Vaughn [212] is 1.23 Gb, approximating more closely the 0.93 Gb *Artemia* GS estimated in this study.

Currently, out of the 50,000 known crustacean species, the GS of 278 crustaceans has been determined, covering a 400-fold-wide genome size range between *Cyclops kolensis*, a cyclopoid copepod (0.14 pg) and *Ampelisca macrocephala*, an Arctic amphipod (64.62 pg) [73,88]**.** In comparison, *A. franciscana* has a relatively small genome of 0.97 pg. This makes it a potential new model crustacean for which genome sequencing is currently feasible, unlike for crustaceans with a much larger genome size. To date, the only annotated crustacean genome is that of the branchiopod *D. pulex*, with an average genome size of 0.23 pg [217].

Ultimately, the further development of genomic resources for *Artemia* such as the whole-genome sequence, will add a completely new dimension to *Artemia* research.

Moreover, knowledge of the *A. franciscana* sex-determining system will facilitate future evolutionary studies of sex chromosomes in sexually dimorphic (WZ female/ZZ male) and parthenogenetic *Artemia.* Considering the presence of sexual and asexual reproduction strategies, the *Artemia* genus shows promise as a model system for the study of asexuality, its evolution and its evolutionary purpose. Finally, since *Artemia* is considered a potential crustacean model species, increasing knowledge about *Artemia* genetics and genomics in general and sex-related genetics in particular, are expected to be valuable to crustacean aquaculture, presently lacking in molecular breeding strategies despite their contribution of 23% to the total aquaculture production value [17]. Future research is discussed in Chapter 5.

Chapter 3

Further characterization of the sex-determining region in *Artemia*

3.1. Abstract

Previously, by linkage mapping, we have identified the sex-determining region as a single locus segregating in the *Artemia* genome, disclosing the female as the heterogametic sex (Chapter 2). Here, bulked segregant analysis (BSA) by high-throughput sequencing of *Artemia franciscana* with subsequent selection revealed eight candidate primary sex-determining genes in the *Artemia* genome. Four candidate genes contained exonic non-synonymous SNPs that affected secondary protein structure: *Cytochrome P450*, *F0F1 ATP synthase subunit beta*, a gene containing a CRAL-TRIO domain and a gene containing an ankyrin repeat. Two candidate genes were *Fibronectin* genes with exonic non-synonymous SNPs that had no effect on secondary protein structure. Two final candidate genes were *SEC14* and *Zinc finger C2CH-type,* which contained no exonic SNPs.

Several genes, highly homologous with insect sex-determining genes (*doublesex*, *sex-determining region Y*, *sex-lethal*, *feminization 1*, *transformer 1*, *transformer 2*, *fruitless*, *runt*, *deadpan*, *daughterless*, *extra macrochaetae*, *groucho* and *sans fille*) or with crustacean sex-related genes (*extra macrochaetae*, *fushi tarazu*, *forkhead box*, *WNT*, *argonaute*, *testis-specific*, *VASA*, *SOX*, *star*, *ECM*, *tudor*, *GATA*, *prohibitin* and *Cytochrome P450*) were present in the *Artemia* genome as well. Of these arthropod sex-related genes, only *Cytochrome P450* was identified as part of the sex-determining region by the BSA analysis and showed a sex-specific secondary protein structure, indicating that it may play a role in primary sex determination in *Artemia. Cytochrome P450* represented the most valid candidate sex-determining gene, because it has already been put forward as a candidate sex-determining gene in crustacean *M. nipponense* through transcriptomic evidence.

3.2. Introduction

3.2.1. Importance of sex-determining genes

Crustacean aquaculture is an over US$28 billion food production industry [56]. As most cultured species have a sexually dimorphic growth [215], yield in crustacean aquaculture would be greatly improved by identification of the primary sex-determining genes, enabling development of molecular breeding strategies and RNA interference (RNAi) technology approaches to achieve durable monosex cultures of the faster-growing sex [215]. Identified upstream sex-determining genes in *Artemia* might facilitate orthologous gene discovery in other farmed crustaceans [129]. This research will help understanding sex determination mechanisms in crustaceans, as well as enable the design of new genetic sexing strategies and possibly monosex breeding in different crustacean species of economic importance.

3.2.2. Sex determination mechanisms in arthropods

3.2.2.1. Insects

Sex determination mechanisms in arthropods (section 1.3.) have been elucidated in great detail in the model insect *Drosophila melanogaster* (Figure 1-12). In insects, sex is primarily determined by chromosomal signals (genetic sex determination; Figure 1-10) [218].

Several upstream sex-determining genes (*sxl*, *tra, tra-2, fem* and *csd*) have been identified in different insect species, but are still unidentified in WZ/ZZ sex determination systems (section 1.3.). Downstream sex-determining genes in insects include *dsx* and *fru*. The farther upstream in the sex determination cascade, the less conserved the sex-determining genes are (e.g. *sxl* and *tra*), whereas downstream sex-determining genes, such as *dsx* and *fru* are highly conserved in a broad range of insects and even in crustaceans [33,63,97]

3.2.2.2. Crustaceans

Cladocerans

Research on sex determination mechanisms in *Daphnia magna*, a freshwater branchiopod crustacean, gave rise to the concept of environmental sex determination (ESD) in arthropods, for which it became a model organism [97]. *D. magna* parthenogenetically produces males in response to environmental cues.

Two *doublesex* (*dsx*) paralogs, *DapmaDsx1* and *DapmaDsx2* are expressed in *D. magna* males only. Expression of *DapmaDsx1* is responsible for male trait development during ESD and shows similar domains to the insect *dsx* [66,97]. The specific function of *DapmaDsx2* remains unknown [97]. In the cladocerans *D. pulex*, *D. galeata*, and *Ceriodaphnia dubia*, homologs of both *DapmaDsx* genes showing male-biased expression were identified*,* but only one single *dsx* gene was found in *Moina macrocopa* [202]. A *tra* ortholog has been characterized in *D. magna* as well, however it shows no sexual dimorphic differences in expression or splicing, and, hence, is not responsible for sex determination in *D. magna* [96]. A *csd*-like protein has been found in *D. pulex* [39].

Currently in crustaceans, mainly highly conserved downstream sex-determining genes such as *dsx* and (even more downstream) hormones, such as androgenic gland-specific insulin-like peptides (IAGs) [215] have been identified, the latter allowing hormonally induced sex reversal. In *D. magna*, male production occurs independently of environmental cues by treatment with the exogenous juvenile hormone (JH) or its analogs [202].

Shrimps and prawns

Several projects have aimed at identifying sex differentiation genes in shrimps and prawns, mainly by expression analysis in gonads or during gonad development [90,101,110,124,125,163], but no primary sex-determining gene has been identified.

A *tra-2* homolog (*FcTra-2*) has been cloned and characterized in the Chinese shrimp *Fenneropenaeus chinensis*, revealing three splice isoforms (*FcTra-2a*, *FcTra-2b and FcTra-2c*) [117]. Expression of the isoform *FcTra-2c* suddenly increases at the mysis stage, is significantly higher in juvenile females than in males and is significantly higher in ovary than in other tissues, all suggesting that *FcTra-2* might be involved in *F. chinensis* sex determination [117]. In the tiger shrimp *Penaeus monodon,* the *dsx* gene is rooted ancestrally to both DSX1 and DSX2 of cladocerans (Figure 1-13) [202].

Based on transcriptome analysis of the oriental river prawn, *Macrobrachium nipponense*, putative genes involved in crustacean sex determination and sex differentiation were identified, including *sex-determining region Y-chromosome* (*SRY*), *doublesex*- and *mab-3-Related transcription factor 1* (*DMRT1*), *forkhead box L2* (*FOXL2*), *feminization 1* (*FEM1*), *fushi tarazu factor 1* (*FTZ-F1*), *VASA* and other potential candidates [124].

They are homologous to known sex determination and sex-differentiation genes, but, currently, no experimental evidence is available to support their function.

Salmon louse

Besides *Daphnia*, salmon louse (copepod *Lepeophtheirus salmonis*) is the only other crustacean species of which a genome sequence is available [85]. L. *salmonis* has a female heterogametic system and a sex-linked SNP has been located in the coding region of *prohibitin-2*, showing sex-dependent differential expression with mRNA levels 1.8-fold higher in adult females than males [27].

Artemia

A *dsx*-related fragment from the parthenogenetic *Artemia* (GenBank: AY346101.1) and a DM domain contained in human and mouse *Dmrt3* (83% identity) from *A. sinica* and parthenogenetic *Artemia* cDNA have been submitted [232]. DM domain genes, such as *dsx* in insects and *Dmrt* in vertebrates, determine sexual development and its evolution in many metazoans [137].

In Chapter 2, the sex determining region has been identified by linkage mapping as a single locus segregating in the *Artemia* genome, disclosing the female as the heterogametic sex. In agreement with classical ancient, highly evolved sex chromosomes, each of the two homologous sex linkage groups show a relatively small recombining pseudo-autosomal region (PAR) and large non-recombining regions [18].

Here, we identified putative sex-determining genes by means of bulked segregant (BSA) analysis by high-throughput sequencing.

3.3. Materials and methods

3.3.1. Rationale for potential *Artemia* sex determination systems

A. franciscana is known to have heteromorphic sex chromosomes (Chapter 1) and, as shown in Chapter 2, females are the heterogametic sex (WZ/ZZ sex-determining system). We propose the following three hypothetical WZ/ZZ sex-determining systems:

-1) The primary sex-determining gene lies on the Z chromosome

This situation is depicted in Figure 3-1.

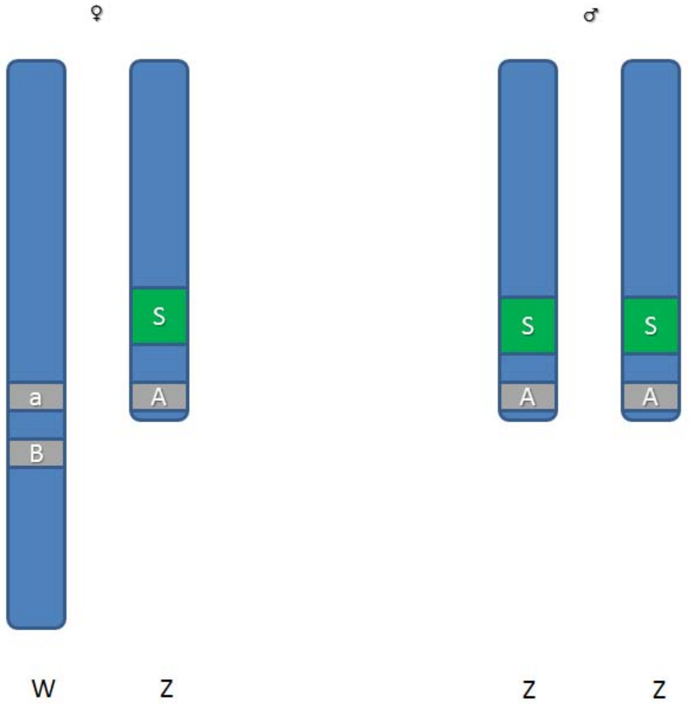


Figure 3-1. A WZ/ZZ sex-determining system showing the the sex-determining gene present on the Z chromosome (green, S). Chromosomes W and Z are heteromorphic (blue). The A and B allele represent the "AFLP band present" allele and are in linkage disequilibrium with the S-allele of the sex locus.

This hypothesis implies that the sex-determining gene is present in both the male and female genome. Such a system occurs in chicken, in which sex is determined by dosage of the sex-determining gene *dmrt1*, present on the Z chromosome [189]. AFLP analysis of such a WZ/ZZ dosage-dependent sex-determining system will result in S(Z)-linked AFLP markers segregating as homozygous present in males and heterozygous in females, analogous to pattern 3 in Table 2-3, or in W-linked AFLP markers, segregating as present in females and absent in males, analogous to pattern 1 described in Table 2-3.

-2) The primary sex-determining gene lies on the W chromosome

This situation is depicted in Figure 3-2.

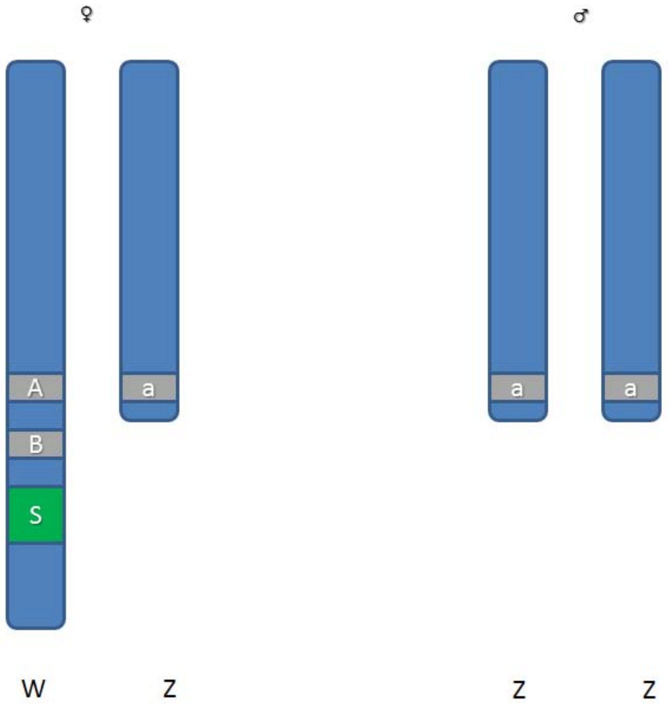


Figure 3-2. A WZ/ZZ sex-determining system showing the sex-determining gene present on the W chromosome (green, S). Chromosomes W and Z are heteromorphic (blue). The A and B allele represent the "AFLP band present" allele and are in linkage disequilibrium with the S-allele of the sex locus.

This hypothesis implies that the sex-determining gene is present only in the female genome. This system occurs in *Bombyx mori*, in which sex is determined by one gene, present on the W chromosome. AFLP analysis of such a WZ/ZZ sex-determining system will result in S-linked AFLP markers, segregating as present in females and absent in males, analogous to pattern 1 described in Table 2-3.

-3) The primary sex-determining gene lies on both the W and Z chromosomes.

This situation is depicted in Figure 3-3:

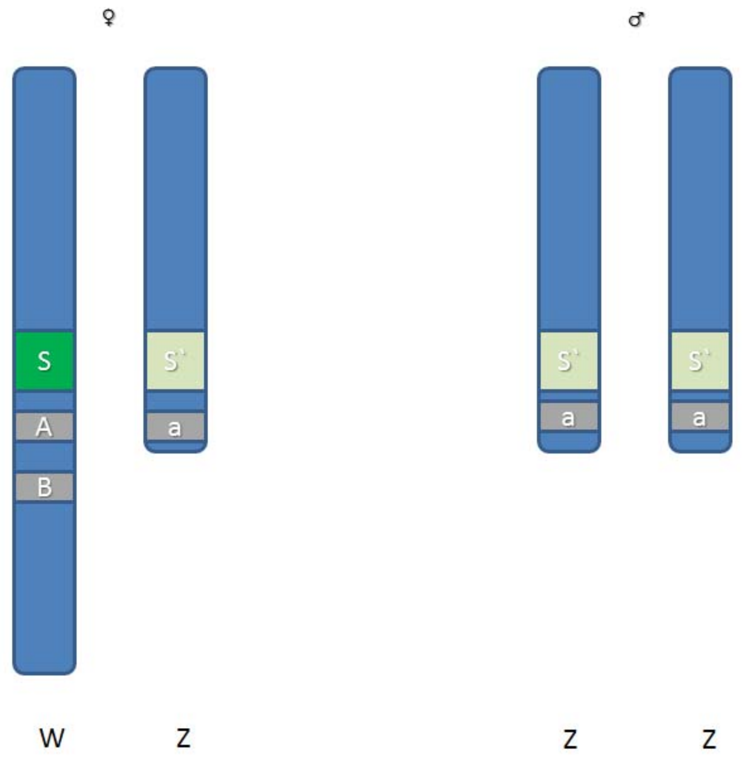


Figure 3-3. A WZ/ZZ sex-determining system showing the sex-determining gene present on both the W and Z chromosomes as allelic variants (green, S and light green, S'). Chromosomes W and Z are heteromorphic (blue). The A and B allele represent the "AFLP band present" allele and are in linkage disequilibrium with the S-allele of the sex locus.

This hypothesis implies allelic variants at the sex-determining locus. Such a system occurs, for example, in tiger pufferfish *Takifugu rubripes* (XY/XX), in which sex is determined by a single SNP in gene *Amhr2* [94]. AFLP analysis of a WZ/ZZ sex-determining system as described in Figure 3-3 will result in S-linked AFLP markers, segregating as homozygous absent in males and heterozygous in females, analogous to pattern 1 described in Table 2-3.

The three hypothetical WZ/ZZ sex-determining systems proposed in section 3.3.1. are possible in *Artemia*: a primary sex-determining gene lying on the Z chromosome (Table 2-3, patterns 1 or 3), on the W chromosome (Table 2-3, pattern 1), or on both the W and the Z chromosomes (Table 2-3, pattern 1), because sex-linked AFLP markers following pattern 1 were found for *Artemia*.

As all sex-linked AFLP markers found (Chapter 2) were heterozygous in females and homozygous in males, generating markers for which females are heterozygous and males are homozygous should enable us to identify candidate sex-determining genes.

3.3.2. AFLP analysis, sequencing and BLAST of sex-linked markers

The seven sex-linked AFLP marker fragments estimated longer than 100 bp by AFLP-Quantar*Pro* (E112M341M405.3F; E112M223M273.0F; E112M232M351.5F; E112M332M350.8F; E112M134M244.0F; E112M133M245.4F and E112M142M210.7F; Chapter 2) were purified from polyacrylamide gels, followed by PCR amplification and subsequent Sanger sequencing as described in Vuylsteke *et al*. [224]. Then, nucleotide-nucleotide BLAST (blastn, (Expectation value 1E-10) of these sex-linked markers was carried out on the draft *Artemia* genome assembly (Chapter 4).

3.3.3. BSA by whole-genome sequencing

A bulked segregant analysis strategy (BSA; section 1.2.1.2) was developed to detect SNP markers for which females are heterozygous and males are homozygous.

Cyst material of *A. franciscana* strains from San Francisco Bay (SFB) and Vinh Chau (VC) was hatched and reared until sexual maturation as described in Chapter 2. A first controlled cross (C1) between VC (♀) and SFB (♂) was made, resulting in F_1 full-sib progeny that was collected, grown until maturity, rinsed and sexed (Figure 1.3.). For gut evacuation before DNA extraction, the offspring were kept overnight in a cellulose solution (1.5 g/l; Sigma, type 20) [41]. The brood pouch was removed in females. Progeny were stored individually at -20 °C.

Briefly, BSA was done by (1) mapping reads of a male and a female bulk onto the *de novo* assembled *Artemia* genome (Chapter 4); (2) identifying of SNPs; and (3) selecting of scaffolds containing SNPs with significant differences in allele frequency between female and male bulk reads.

For paired-end (PE) sequencing, DNA was extracted from each of the 120 F_1 progeny obtained from C1. DNA quality and yield were assessed by NanoDrop ND 1000 spectrophotometry (Thermo Scientific) [48] and samples were diluted to a DNA concentration of 30 ng/µl. Fifty-five equimolar male F_1 progeny and 65 equimolar female F_1 progeny DNA samples were pooled.

To improve assembly results, Illumina DNA sequencing libraries of two different insert sizes (200 and 500 bp) were prepared from each pool (1-5 µg DNA) with the standard protocol for the "Illumina TruSeq DNA Sample Preparation Kit" and were sequenced by Fasteris (http://www.fasteris.com). DNA fragmentation (50-500 bp), end-repair, A-tailing and adapter ligation were followed by gel isolation of fragments of the required insert size and PCR amplification. The libraries were quantified, diluted to 10 nM and analyzed in a paired-end 100-bp run on an Illumina HiSeq 2000 instrument.

PE reads from the male and female pools were each aligned to the *Artemia* draft genome (Chapter 4) by means of the Burrows-Wheeler Alignment tool[4] [113]. Mapping files were processed with Samtools[5] [114] to filtered pileup files representing only uniquely aligned reads. SNPs were detected with the “somatic option” in VarScan[6] [105] (hereafter referred to as “detected SNPs”). In somatic mode, VarScan reads two pileup files simultaneously. Positions that are present in both files, and meet the minimum coverage (8X) in each are compared. Whereas allele frequencies in the male and female pools should be approximately equal in all genomic regions except loci affecting sex, they should differ at the genomic region containing the sex-determining gene(s)[34]. Consistent with a single sex-determining gene model in which *Artemia* females are the heterogametic sex, SNP loci exhibiting heterozygosity in the female pool reads and homozygosity in the male pool reads are potentially valid sex-linked SNPs (section 3.3.1.).

---

[4] BWA version 0.5.9

BWA commands per paired-end read library:

```
bwa index genome.fasta
bwa aln –t5
bwa sampe
```

BWA commands for single-end reads from a library, created by read trimming of paired-end reads:

```
bwa index genome.fasta
bwa aln –t5
bwa samse
```

[5] Samtools version 0.1.19

Samtools commands per read library:

```
samtools faidx genome.fasta
samtools view –bS
samtools sort
samtools rmdup
samtools mpileup –q1
```

parameter -q1 filters for uniquely aligned reads

[6] VarScan version 2.3.3

Varscan commands for male and female vcf files:

```
VarScan somatic female.vcf male.vcf varscanout --validation=1
```

Only SNPs with a SNP coverage of ≥ 15X per pool [50], with homozygous allele frequencies in male reads[7] and heterozygous allele frequencies in female reads[8], present on scaffolds carrying five or more SNPs, were considered for further analysis (hereafter referred to as “selected SNPs”).

The risk of identifying sequencing errors as “heterozygous” loci was minimized in several ways. A quality trim of the NGS reads (Chapter 4) kept only base calls with a minimal quality score Q of 20, which have a typical error rate of 1%. Also, the minimal SNP coverage of ≥ 15X per pool and selection of scaffolds containing a minimum of five selected SNPs minimizes the possibility that one sequencing error might be perceived as a SNP.

Cut-off values for calling homo- and heterozygosity were determined as follows. Assume a SNP with two alleles “A” and “T”, with respective proportions *p* and 1-*p* in a theoretical population of all possible reads in a diploid species. As a consequence, random drawing of a read can be considered as a Bernoulli event with probability *p*. For a coverage of *N* reads, considered as *N* independent Bernoulli events following the binomial distribution, the probability of having *k* times allele “A” can be calculated as:

$$\boldsymbol{P\,(X = k) = \binom{N}{k}\, p^{k}\,(1 - p)^{\,N-k}}$$

So far, we did not assume any sequencing error. Consider sequencing errors happen with a probability *t*. Hence, an “A” allele will be detected as a true “A” with probability *p(1-t)* and as a false “A” with probability (1-*p*)*t*/3. Assuming equal error rates for all nucleotides, a “T” may be erroneously read as an “A”,”C” or “G”, each option with an equal probability of t/3. Hence, for a coverage of *N* reads, the probability of having *k* times allele A becomes:

$$\boldsymbol{P(X = k) = \tbinom{N}{k}\; q^{k}(1 - q)^{N-k}}$$

with $q = p\,(1-t) + (1-p)\,\frac{t}{3}$

In order to find appropriate cut-off values for calling homo-and heterozygosity, the threshold *S* for which the probability for homozygosity (P_1) is equal to that of heterozygosity (P_2) is determined:

$$\boldsymbol{P_1 = P_2}$$

---

[7] $A_1 \geq 0.90(A_1+A_2)$ or $A_2 \geq 0.90(A_1+A_2)$
A_1 and A_2 represent the allele frequencies of the reference and the variant base, respectively

[8] $0.25(A_1+A_2) \leq A_1 \leq 0.75(A_1+A_2)$ or $0.25(A_1+A_2) \leq A_2 \leq 0.75(A_1+A_2)$

$$P_1(X=k)=\binom{N}{k}\ q_1^k(1-\ q_1)^{N-k}\ =\ P_2(X=k)=\binom{N}{k}\ q_2^k(1-\ q_2)^{N-k}$$

With $q_1=p_1(1-t)+(1-p_1)\frac{t}{3}$

and $q_2=p_2(1-t)+(1-p_2)\ \frac{t}{3}$

$$<=>S=\frac{k}{N}=\frac{\log\left(\frac{1-q_2}{1-q_1}\right)}{\log\left(\frac{q_1}{q_2}\right)+\log\left(\frac{1-q_2}{1-q_1}\right)}$$

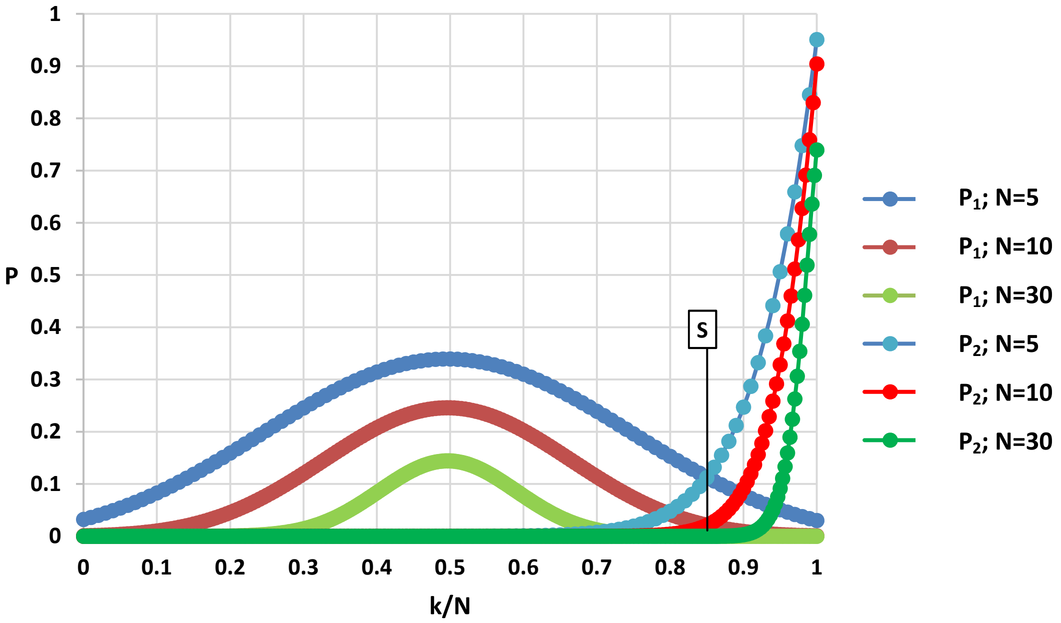


Figure 3-4. Assuming a SNP covered N times with alleles "A" and "T", distribution of the probability P of having k times an allele "A". P is shown for the cases of heterozygosity (P_1) and homozygosity (P_2). S=0.85 is the threshold where P_1=P_2, in other words, where the probability of heterozygosity is equal to that of homozygosity.

It is clear from the above mentioned formula for S and from Figure 3-4 that S depends on t and not on N. S is the same for any coverage, but, owing to the binomial distribution of P, the risk of misconclusion decreases with increasing coverage.

Assuming a typical Illumina platform error rate of 1%, S equals 85% (Figure 3-4). Based on a threshold S of 85% where homozygosity could not be distinguished from homozygosity, instead of only omitting results for SNPs with the exact allele frequency of 85%, to be sure, we took an arbitrary allele frequency interval]75%; 90%[containing S, where SNPs were not analyzed. Therefore, the lower cut-off for homozygosity calling was set at 90%, and the upper cut-off for heterozygosity calling was set at 75%.

A linear mixed model, including splines, initially proposed by Claesen *et al*. [34] and implemented in this study with the multifunctional statistical tool Genstat 15 [86], was fitted to the logit-transformed binomial data to discover scaffolds that potentially contain the sex-determining gene. False discovery rate (FDR) was estimated by modeling significance values as a 2-component mixture of Uniform and Beta or Gamma densities [10], as implemented in Genstat. Scaffold probabilities were computed by calculating the mean per scaffold of the binomial probabilities for the actual male and female allele frequency per SNP, given that expected allele frequencies are 50:50 in females and 100:0 in males. Scaffolds with FDR < 0.01 and a scaffold probability > 0.5 were selected as candidate scaffolds holding the sex locus (hereafter referred to as "selected scaffolds"). Genes lying on selected scaffolds, holding selected SNPs within the gene region, were selected when they held a function linked with sex determination, based on the literature (hereafter referred to as "selected genes"). Spline plots were generated for scaffolds containing sex-linked AFLP markers or selected genes.

3.3.4. Sex-specific genes with non-synonymous SNPs

One way of assessing a putatively sex-linked gene is to test whether SNPs from male and female reads, found in the gene, are synonymous or not. Synonymous SNPs have different alleles that encode for the same amino acid, while non-synonymous SNPs have different alleles that encode for different amino acids. In the described experimental setup, genes carrying non-synonymous SNPs are expected to produce a sex-specific protein. Selected genes showing heterozygosity in females and homozygosity in males on spline plots [34] of detected and selected SNPs [34] were analyzed for sex-specificity by looking for non-synonymous SNPs.

In selected exonic SNPs of selected genes, "reference" bases of the gene were altered to "variant" bases, the coding DNA sequence (CDS) of the "variant" gene was annotated and translated[9] to its corresponding "variant" protein and the "reference" protein and "variant" protein were aligned[10]. "Variant" proteins showing amino acids, different from their respective reference protein were termed genes containing non-synonymous SNPs. In order to evaluate the effect of non-synonymous SNPs in variant proteins, the secondary protein structure was predicted for each non-synonymous SNP separately[11].

---

[9] Standard genetic code translation table

[10] Gap open cost 10.0; Gap extension cost 1.0, CLC Main Workbench 6.7.1.

[11] CLC Main Workbench 6.7.1

3.3.5 Arthropod sex-determining gene homologs, not selected by BSA

All the arthropod sex-determining genes mentioned in sections 1.3 and 3.2 were searched by browsing through the complete *Artemia* genome on the Orcae online genome annotation platform (Chapter 4) with the “keyword” search function.

3.4. Results

3.4.1. AFLP analysis, sequencing and BLAST of sex-linked markers

Scaffold_13 was produced by anchoring scaffolds, each holding one or two sex-linked AFLP markers in the order indicated by the *Artemia* high density genetic linkage map (Chapter 2). Orientation data was only applicable to scaffolds containing more than one sex marker. Anchoring was detailed in a golden path (AGP) file (Table 3-1).

Table 3-1. AGP file for LG_Female_1, the new scaffold created by anchoring of SM-containing scaffolds (in their turn made out of contigs), according to the *Artemia* genetic linkage map. Scaffold12172_size22368_rc stands for the reverse complement of scaffold12172_size22368. For each contig is provided: the beginning (beg) and end (end) positions in LG_Female_1; its number (nr), type (W = WGS contig, N = gap with specified size); name (component_id); the beginning (component_beg) and end positions (component_end) within the contig and the orientation of the contig within LG_Female_1. For each gap is provided: gap length (gap_length), the gap type (gap_type, in this case, all gaps are due to scaffolding or anchoring), the presence of linkage (linkage) and the linkage evidence (linkage_evidence).

object	beg	end	nr	type	component_id /gap_length	component_beg /gap_type	component_end /linkage	Orientation /linkage_evidence
LG_Female_1	1	993	1	W	scaffold21861_size15620_contig_1	1	993	+
LG_Female_1	994	1014	2	N	21	scaffold	yes	paired-ends
LG_Female_1	1015	5133	3	W	scaffold21861_size15620_contig_2	1	4119	+
LG_Female_1	5134	5174	4	N	41	scaffold	yes	paired-ends
LG_Female_1	5175	7991	5	W	scaffold21861_size15620_contig_3	1	2817	+
LG_Female_1	7992	8032	6	N	41	scaffold	yes	paired-ends
LG_Female_1	8033	9275	7	W	scaffold21861_size15620_contig_4	1	1243	+
LG_Female_1	9276	9286	8	N	11	scaffold	yes	paired-ends
LG_Female_1	9287	15702	9	W	scaffold21861_size15620_contig_5	1	6416	+
LG_Female_1	15703	20702	10	N	5000	anchor	yes	map
LG_Female_1	20703	20809	11	W	scaffold17845_size17924_contig_6	1	107	+
LG_Female_1	20811	22518	12	N	1600	scaffold	yes	paired-ends
LG_Female_1	22519	25122	13	W	scaffold17845_size17924_contig_7	1	2604	+
LG_Female_1	25123	25163	14	N	41	scaffold	yes	paired-ends
LG_Female_1	25164	34617	15	W	scaffold17845_size17924_contig_8	1	9454	+
LG_Female_1	34618	35657	16	N	1040	scaffold	yes	paired-ends
LG_Female_1	35658	35793	17	W	scaffold17845_size17924_contig_9	1	136	+
LG_Female_1	35794	35834	18	N	41	scaffold	yes	paired-ends
LG_Female_1	35835	38742	19	W	scaffold17845_size17924_contig_10	1	2908	+
LG_Female_1	38743	43742	20	N	5000	anchor	yes	map
LG_Female_1	43743	44284	21	W	scaffold12172_size22368_rc_contig_11	1	542	+
LG_Female_1	44285	44388	22	N	104	scaffold	yes	paired-ends
LG_Female_1	44389	46116	23	W	scaffold12172_size22368_rc_contig_12	1	1728	+
LG_Female_1	46117	47160	24	N	1044	scaffold	yes	paired-ends
LG_Female_1	47161	51588	25	W	scaffold12172_size22368_rc_contig_13	1	4428	+
LG_Female_1	51589	51700	26	N	112	scaffold	yes	paired-ends
LG_Female_1	51701	53655	27	W	scaffold12172_size22368_rc_contig_14	1	1955	+
LG_Female_1	53656	53697	28	N	42	scaffold	yes	paired-ends
LG_Female_1	53698	60757	29	W	scaffold12172_size22368_rc_contig_15	1	7060	+

object	beg	end	nr	type	component_id /gap_length	component_beg /gap_type	component_end /linkage	Orientation /linkage_evidence
LG_Female_1	60758	60798	30	N	41	scaffold	yes	paired-ends
LG_Female_1	60799	61111	31	W	scaffold12172_size22368_rc_contig_16	1	313	+
LG_Female_1	61112	61234	32	N	123	scaffold	yes	paired-ends
LG_Female_1	61235	63092	33	W	scaffold12172_size22368_rc_contig_17	1	1858	+
LG_Female_1	63093	63271	34	N	179	scaffold	yes	paired-ends
LG_Female_1	63272	64059	35	W	scaffold12172_size22368_rc_contig_18	1	788	+
LG_Female_1	64060	65815	36	N	1756	scaffold	yes	paired-ends
LG_Female_1	65816	66043	37	W	scaffold12172_size22368_rc_contig_19	1	228	+
LG_Female_1	66044	71043	38	N	5000	anchor	yes	map
LG_Female_1	71044	75664	39	W	scaffold4411_size35259_contig_20	1	4621	+
LG_Female_1	75665	75705	40	N	41	scaffold	yes	paired-ends
LG_Female_1	75706	76599	41	W	scaffold4411_size35259_contig_21	1	894	+
LG_Female_1	76600	76847	42	N	248	scaffold	yes	paired-ends
LG_Female_1	76848	93403	43	W	scaffold4411_size35259_contig_22	1	1655	+
LG_Female_1	93404	93449	44	N	46	scaffold	yes	paired-ends
LG_Female_1	93450	95028	45	W	scaffold4411_size35259_contig_23	1	1579	+
LG_Female_1	95029	95049	46	N	21	scaffold	yes	paired-ends
LG_Female_1	95050	95389	47	W	scaffold4411_size35259_contig_24	1	340	+
LG_Female_1	95390	95517	48	N	128	scaffold	yes	paired-ends
LG_Female_1	95518	96311	49	W	scaffold4411_size35259_contig_25	1	794	+
LG_Female_1	96312	96521	50	N	210	scaffold	yes	paired-ends
LG_Female_1	96522	97823	51	W	scaffold4411_size35259_contig_26	1	1302	+
LG_Female_1	97824	97939	52	N	116	scaffold	yes	paired-ends
LG_Female_1	97940	99304	53	W	scaffold4411_size35259_contig_27	1	1365	+
LG_Female_1	99305	99585	54	N	281	scaffold	yes	paired-ends
LG_Female_1	99586	100154	55	W	scaffold4411_size35259_contig_28	1	569	+
LG_Female_1	100155	100195	56	N	41	scaffold	yes	paired-ends
LG_Female_1	100196	100435	57	W	scaffold4411_size35259_contig_29	1	240	+
LG_Female_1	100436	100466	58	N	31	scaffold	yes	paired-ends
LG_Female_1	100467	100580	59	W	scaffold4411_size35259_contig_30	1	114	+
LG_Female_1	100581	100866	60	N	286	scaffold	yes	paired-ends
LG_Female_1	100867	102849	61	W	scaffold4411_size35259_contig_31	1	19831	+
LG_Female_1	102850	102992	62	N	143	scaffold	yes	paired-ends
LG_Female_1	102993	103804	63	W	scaffold4411_size35259_contig_32	1	812	+
LG_Female_1	103805	104058	64	N	254	scaffold	yes	paired-ends
LG_Female_1	104059	106214	65	W	scaffold4411_size35259_contig_33	1	2156	+
LG_Female_1	106215	111214	66	N	5000	anchor	yes	map
LG_Female_1	111215	111443	67	W	scaffold38839_size9725_contig_34	1	229	+
LG_Female_1	111444	111598	68	N	155	scaffold	yes	paired-ends
LG_Female_1	111599	116906	69	W	scaffold38839_size9725_contig_35	1	5308	+
LG_Female_1	116907	116927	70	N	21	scaffold	yes	paired-ends
LG_Female_1	116928	121282	71	W	scaffold38839_size9725_contig_36	1	4355	+

Six of the seven sex-linked AFLP markers aligned significantly with the draft genome on scaffold_13 (Table 3-2), using BLAST (blastn, E-value 1E-10).

Table 3-2. BLAST results of six sex-linked AFLP markers on the *Artemia* genome (Scaffold_13). Female_1 is the female linkage group on which the sex-linked markers are located in the *Artemia* linkage map (Figure 2-4).

Name AFLP fragment	Length AFLP fragment (bp)	Position AFLP fragment on Female_1 (cM)	Hit start on scaffold_13	Hit end on scaffold_13	ID	E-value
E112M341M405.3F	331	5.197	1015	1178	100.00	1.00E-80
E112M223M273.0F	216	5.270	31640	31844	95.65	2.00E-87
E112M232M351.5F	281	5.274	52500	52766	100.00	6.00E-138
E112M332M350.8F	294	5.308	52545	52825	98.94	1.00E-140
E112M133M245.4F	103	5.315	89062	89147	98.84	3.00E-36
E112M142M210.7F	87	5.402	118032	118115	88.24	4.00E-20

Scaffold_13 (Figure 3-5), was selected by Genstat and had a mean binomial probability > 0.5. Genes present on Scaffold_13 were not selected because they did not contain selected SNPs within the CDS. One selected SNP and two indels were present within the location of one of the sex-linked AFLP markers that aligned with the assembly (E112M223M273.0F) as shown in Figure 3-6. Neither the SNP, nor the indels were located within the restriction sites or selective bases of the marker.

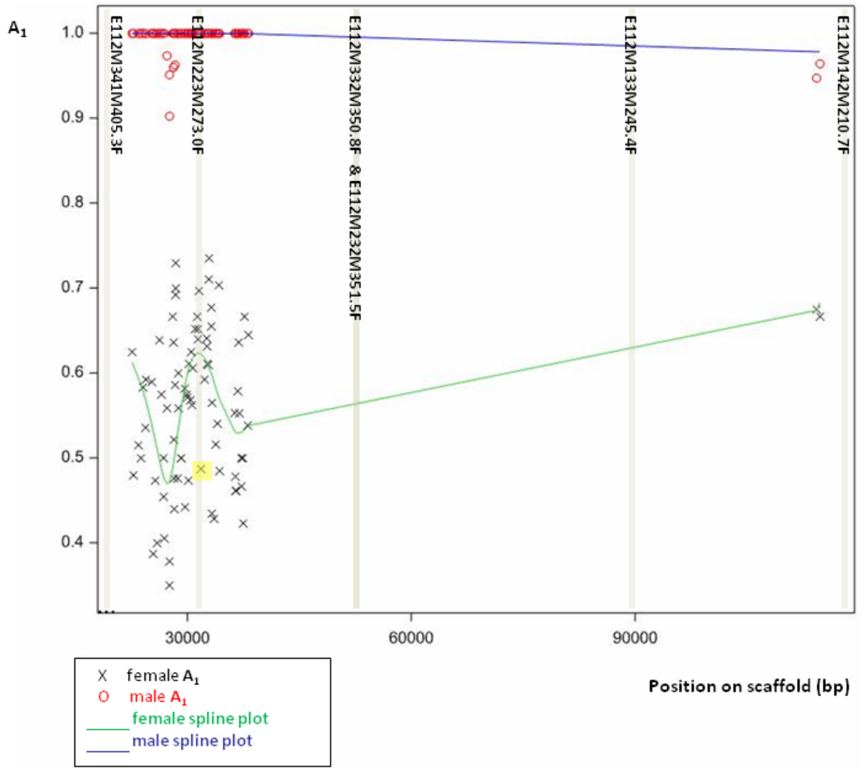


Figure 3-5. Scaffold_13, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Shown in grey, sex-linked AFLP markers aligning to the *Artemia* genome assembly; in yellow, the selected SNP present in marker E112M223M273.0F. Markers E112M232M351.5F and E112M332M350.8F overlap partly. Scaffold_13 was constructed by anchoring scaffolds each holding at least one sex-linked AFLP marker.

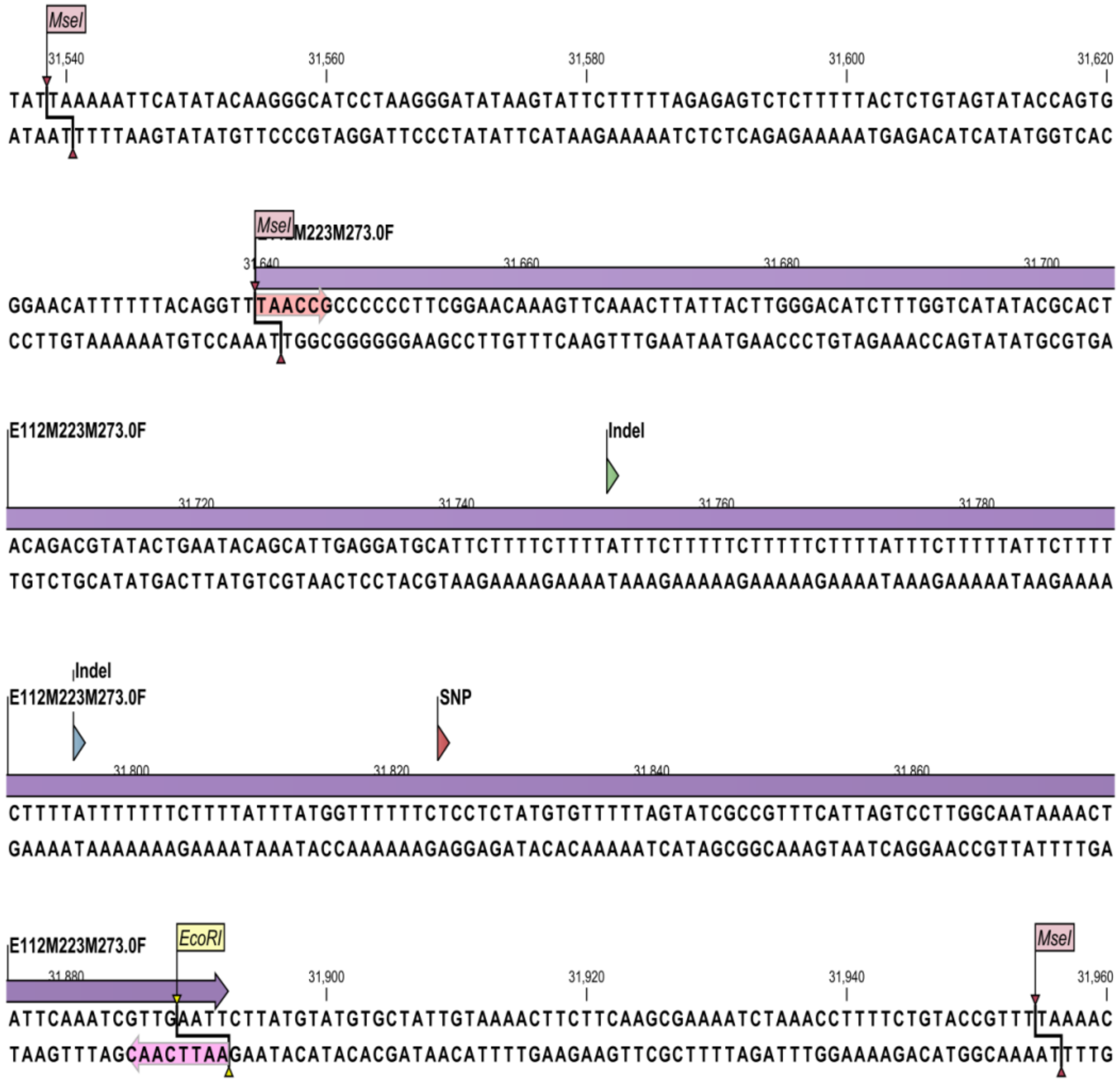

MseI
31,540
31,560
31,580
31,600
31,620
TATTAAAAATTCATATACAAGGGCATCCTAAGGGATATAAGTATTCTTTTTAGAGAGTCTCTTTTTACTCTGTAGTATACCAGTG
ATAATTTTTAAGTATATGTTCCCGTAGGATTCCCTATATTCATAAGAAAAATCTCTCAGAGAAAAATGAGACATCATATGGTCAC
MseI
M223M273.0F
31,640
31,660
31,680
31,700
GGAACATTTTTTACAGGTTTAACCGCCCCCCTTCGGAACAAAGTTCAAACTTATTACTTGGGACATCTTTGGTCATATACGCACT
CCTTGTAAAAAATGTCCAAATTGGCGGGGGGAAGCCTTGTTTCAAGTTTGAATAATGAACCCTGTAGAAACCAGTATATGCGTGA
E112M223M273.0F
Indel
31,720
31,740
31,760
31,780
ACAGACGTATACTGAATACAGCATTGAGGATGCATTCTTTTCTTTTATTTCTTTTTCTTTTTCTTTTATTTCTTTTTATTCTTTT
TGTCTGCATATGACTTATGTCGTAACTCCTACGTAAGAAAAGAAAATAAAGAAAAAGAAAAAGAAAATAAAGAAAAATAAGAAAA
Indel
E112M223M273.0F
SNP
31,800
31,820
31,840
31,860
CTTTTATTTTTTTCTTTTATTTATGGTTTTTTCTCCTCTATGTGTTTTTAGTATCGCCGTTTCATTAGTCCTTGGCAATAAAACT
GAAAATAAAAAAAGAAAATAAATACCAAAAAAGAGGAGATACACAAAAATCATAGCGGCAAAGTAATCAGGAACCGTTATTTTGA
E112M223M273.0F
EcoRI
MseI
31,880
31,900
31,920
31,940
31,960
ATTCAAATCGTTGAATTCTTATGTATGTGCTATTGTAAAACTTCTTCAAGCGAAAATCTAAACCTTTTCTGTACCGTTTTAAAAC
TAAGTTTAGCAACTTAAGAATACATACACGATAACATTTTGAAGAAGTTCGCTTTTAGATTTGGAAAAGACATGGCAAAATTTTG

Figure 3-6. Selected SNPs (red flag) and detected indels (green flag) present in Scaffold_13 at the location of sex-linked AFLP marker E112M223M273.0F (purple). Scaffold_13 is shown double stranded with indicated *Eco*RI (yellow) and *Mse*I (pale pink) restriction sites as well as restriction sites including selective bases (red and pink arrows).

Base coverage of scaffold_13 (Figure 3-7) revealed that parts of the scaffold were covered by neither male nor female reads, representing gaps from scaffolding or anchoring of the original sex-linked-marker-containing scaffolds. Other parts of scaffold_13 were covered by either only female or male reads, implying these parts were sex-specific. For instance three sex-linked AFLP markers were only covered by female reads (E112M341M405.3F; E112M232M351.5F and E112M332M350.8F), whereas the remaining parts of scaffold_13 and the three remaining sex-linked AFLP markers (E112M223M273.0F; E112M133M245.4F; E112M142M210.7F) were covered by both male and female reads. Areas where the coverage was much higher, such as in the area beyond 11 kb, likely represent collapsed repeats.

Scaffold_13 did not equal the W or Z sex chromosomes, but because it contained the sex-linked AFLP markers and was covered by both male and female reads, it clearly was part of both the W and the Z chromosomes. In most assemblies of large genomes (except large labor- and cost-intensive projects, such as the human and chicken genomes), haplotypes remain largely unresolved, as is also the case in the *Artemia* genome.

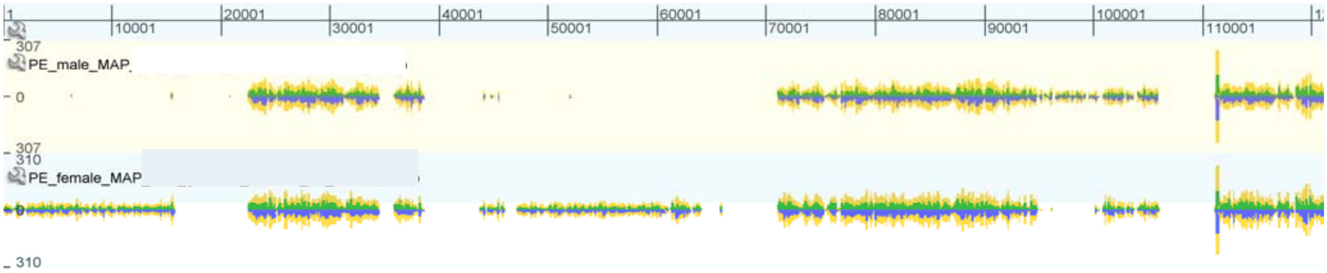


Figure 3-7. Base coverage of scaffold_13 by male (PE_male_MAP) and female (PE_female_MAP) PE reads in pileup format shown in GenomeView [5]. Each read base (A; C; G and T) is represented by a different color. The top scale represents the base pairs of scaffold_13.

3.4.2. Candidate sex-determining genes selected by BSA

Paired-end sequencing generated 814 and 969 million reads of 100 bp for the male and female pools, respectively, resulting in an initial average coverage of 88X and 104X, respectively, based on the 0.93 Gb *Artemia* genome size estimated by flow cytometry (Chapter 2).

VarScan analysis detected 3,882,673 SNP positions and 545,195 indels on the *Artemia* genome. Only SNPs were further analyzed within the scope of this work. 199 selected scaffolds contained 366 different selected genes: 92 genes with and 274 genes without a functional description. Twenty-eight genes holding a functional description were involved in sex determination or differentiation according to the literature, whereas 64 genes (Table 3-3) were not.

Table 3-3. Genes with known function found on BSA-selected fragments selected by BSA, of which the functional description and homologous proteins (not shown) were found unrelated to sex determination or sex differentiation mechanisms, based on the literature.

Gene ID	**Functional description**
artfr19746g00010	*2-isopropylmalate synthase 2*, partial
artfr22976g00010	*4-alpha-glucanotransferase*
artfr15484g00010	*AGAP002342-PB*
artfr133657g00010	*ARL14 effector protein*-like
artfr10787g00020	*ATP-NAD kinase, PpnK-type, all-beta*
artfr10787g00040	*ATP-NAD kinase, PpnK-type, all-beta*
artfr1627g00070	C. briggsae *CBR-MCD-1* protein
artfr12743g00050	*CAAX amino terminal protease*
artfr5789g00010	*coiled-coil domain-containing protein 132-like*
artfr12372g00030	*Component HyfD of membrane-bound*
artfr15143g00040	conserved hypothetical protein
artfr23150g00040	*DEHA2D17204p*
artfr13740g00040	dipeptide-binding protein
artfr12207g00040	*DNA helicase Pif1*
artfr2829g00030	*DNA helicase Pif1*
artfr573g00040	*DNA helicase Pif1*
artfr3068g00070	*Endonuclease/exonuclease/phosphatase*
artfr1862g00030	*Endonuclease/exonuclease/phosphatase*
artfr2043g00030	*Endonuclease/exonuclease/phosphatase*
artfr2043g00110	*endonuclease-reverse transcriptase HmRTE-e01*
artfr18663g00020	*ETC complex I subunit*
artfr24181g00020	*Exonuclease, phage-type/RecB, C-terminal*
artfr12970g00030	*Exonuclease, phage-type/RecB, C-terminal*
artfr12207g00020	*Helitron helicase-like domain*
artfr11142g00010	*hypothetical protein CAOG_06385*
artfr1862g00020	*hypothetical protein CAPTEDRAFT_79080, partial*
artfr13523g00030	*hypothetical protein CGI_10023639, partial*
artfr19676g00010	*hypothetical protein DAPPUDRAFT_206384*
artfr19301g00020	*hypothetical protein DAPPUDRAFT_307817*
artfr4571g00010	*hypothetical protein DAPPUDRAFT_318822*
artfr21047g00060	*hypothetical protein DAPPUDRAFT_329395*
artfr10787g00030	*hypothetical protein LOC100633611, partial*
artfr24181g00050	*hypothetical protein LOTGIDRAFT_164827*
artfr3068g00040	*hypothetical protein NEMVEDRAFT_v1g46811*
artfr22072g00030	*hypothetical protein Rleg2_5684*
artfr9352g00010	*hypothetical protein TcasGA2_TC015058*
artfr165948g00010	*hypothetical protein*
artfr2043g00050	*hypothetical protein*
artfr2043g00140	*hypothetical protein*
artfr10787g00010	*Inorganic polyphosphate/ATP-NAD kinase*, predicted
artfr21607g00020	*melanoma inhibitory activity protein 3*-like
artfr2043g00040	*NAD kinase-like isoform X1*
artfr7854g00030	*PDZ domain*
artfr3068g00060	*Piso0_000169*

Gene ID	Functional description
artfr21047g00020	*P-loop containing nucleoside triphosphate hydrolase*
artfr23412g00040	*Predicted oxidoreductases related to aryl-alcohol dehydrogenases*
artfr10205g00010	*protein canopy homolog 4*-like
artfr12179g00070	*Protein kinase-like domain*
artfr1210g00050	*putative DNA helicase*
artfr12207g00030	*putative DNA helicase*
artfr12179g00060	*putative protein FAM200B*-like
artfr21607g00010	*putative sulfite oxidase subunit YedZ*
artfr13740g00020	*Reverse transcriptase*
artfr1472g00040	*Reverse transcriptase*
artfr22873g00010	*Reverse transcriptase*
artfr2829g00020	*Reverse transcriptase*
artfr3068g00030	*Reverse transcriptase*
artfr4571g00050	*Reverse transcriptase*
artfr16374g00020	*RNA 3'-terminal phosphate cyclase/enolpyruvate transferase, alpha/beta*
artfr1916g00030	*SJCHGC02887 protein*
artfr12961g00020	*Sodium/potassium/calcium exchanger*
artfr3068g00050	uncharacterized protein LOC100887848, partial
artfr5789g00030	*Vacuolar protein sorting-associated protein 54*
artfr23061g00010	*Zinc/iron permease*

After spline plot analysis of the detected SNPs within the 28 selected genes, 12 genes contained mostly or all SNPs heterozygous in females and homozygous in males (Table 3-5 and Figures 3-8 to 3-13), 16 genes (Table 3-4) had less clear SNP allele frequency differences.

Table 3-4. Genes with known function found on fragments selected by BSA, which were rejected based on less clear SNP allele frequency differences in their respective spline plots.

Gene ID	Functional description
artfr2087g00010	Ankyrin repeat
artfr12382g00010	*Asparagine-linked glycosylation protein 1-like*
artfr11572g00020	*Chitin binding domain*
artfr15473g00030	*Concanavalin A-like lectin/glucanases* superfamily
artfr15143g00030	*CRAL-TRIO domain*
artfr2324g00130	*GroEL-like equatorial domain*
artfr13745g00020	*GTP cyclohydrolase I domain*
artfr1862g00010	*Neurotransmitter-gated ion-channel*
artfr2324g00020	*Nucleotide-binding, alpha-beta plait*
artfr11243g00030	*Origin recognition complex, subunit 2*
artfr19301g00010	*Origin recognition complex, subunit 2*
artfr13785g00020	*Protein-tyrosine phosphatase, receptor/non-receptor type*
artfr4146g00020	*SEC14-like protein 5*-like
artfr14904g00020	*zinc finger MYM-type protein 1*-like
artfr23061g00020	*Zinc/iron permease*
artfr5327g00010	*ADAM-TS Spacer 1*

Table 3-5. Genes with functional description linked to sex determination or differentiation in the literature. Four genes (black font) showed mostly SNPs heterozygous in females and homozygous in males, whereas eight genes (blue font) showed all SNPs heterozygous in females and homozygous in males in the detected SNPs. Gene characteristics are provided from the Orcae online annotation platform. The gene ID consists of: "artfr", which stands for *Artemia franciscana*, the scaffold number and "g", followed by the gene number. The start and end positions of the gene on the scaffold, the functional description of the gene and the homologous protein from the best BLAST hit of the NCBI "nr database" with the gene (blastx, a translated nucleotide-protein database BLAST) are provided as well, completed by ID, E-value and organism of the BLAST hit. Scaffolds of genes in blue are represented in spline plots (Figures 3-8 to 3-13).

Gene ID	Start (bp)	End (bp)	Functional Description	Homologous protein	ID	E-value	Organism
artfr15143g00010	3869	6436	CRAL-TRIO domain	*SEC14-like protein 1*	0.81	4.00E-16	*Danaus plexippus*
artfr15143g00020	8733	9701	*F0F1 ATP synthase subunit beta*	*F0F1 ATP synthase subunit beta*	0.25	9.80E-01	*Helicobacter cetorum*
artfr15484g00020	7713	10959	*Fibronectin, type III*	*Tyrosine-protein phosphatase 99A*	0.57	9.00E-21	*Camponotus floridanus*
artfr15484g00030	11064	12004	*Fibronectin, type III*	*s-adenosyl-methyltransferase mraw*	0.40	3.00E-07	*Culex quinquefasciatus*
artfr2087g00020	8929	9441	Ankyrin repeat	putative ankyrin repeat-containing protein	0.46	8.00E-62	*Danaus plexippus*
artfr21607g00030	11675	13830	*Zinc finger, C2CH-type*	_	_	_	_
artfr23412g00030	6091	8162	*cytochrome P450*	*cytochrome P450*	0.25	9.10E-02	*Postia placenta*
artfr4146g00010	5189	5852	*SEC14-like protein 1*	_	_	_	
artfr1916g00020	8616	9861	*Protein kinase*-like domain	*YRK 2*-like	0.80	2.00E-67	*Takifugu rubripes*
artfr1472g00070	44132	50875	*Histone core*	*CAAX prenyl protease 2*	0.51	3.40E-02	*Culex quinquefasciatus*
artfr18348g00010	2346	4721	*Zinc finger, C2H2*	*Transcription factor castor*	0.59	5.00E-39	*Acromyrmex echinatior*
artfr4571g00040	15113	17217	*WD40/YVTN repeat-containing domain*	*GD16950*	0.30	5.50E-01	*Drosophila simulans*

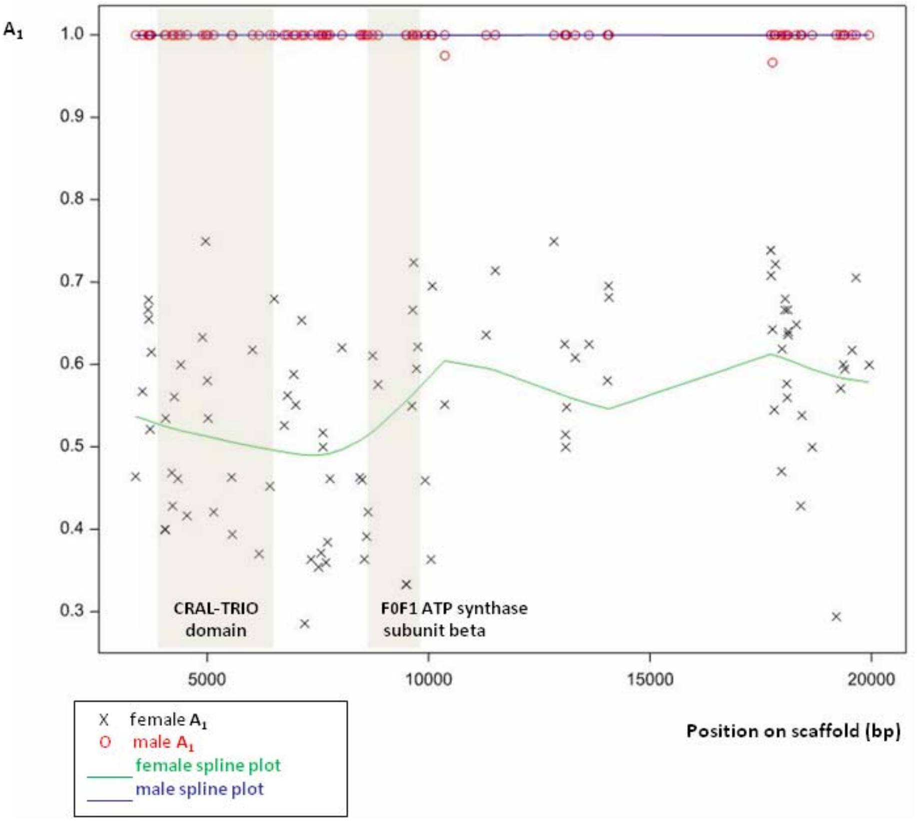


Figure 3-8. Scaffold_15143, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Genes are shown in grey: a gene containing the CRAL-TRIO domain (artfr15143g00010) and *F0F1 ATP synthase subunit beta* (artfr15143g00020).

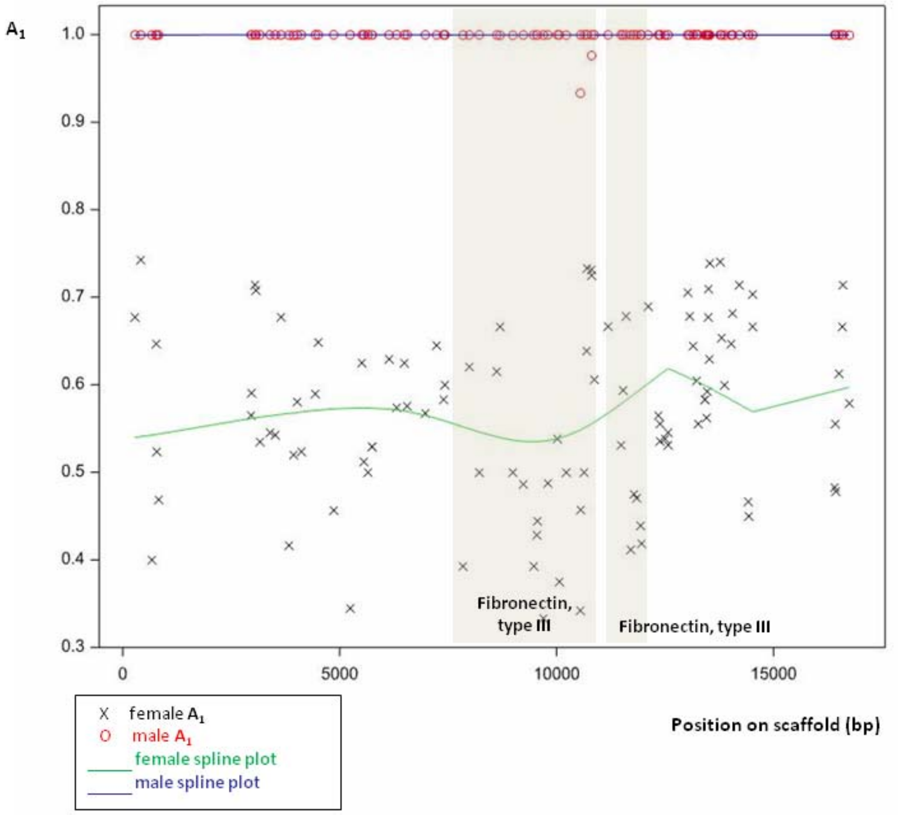


Figure 3-9. Scaffold_15484, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Genes are shown in grey: *Fibronectin, type III* (artfr15484g00020 & artfr15484g00030).

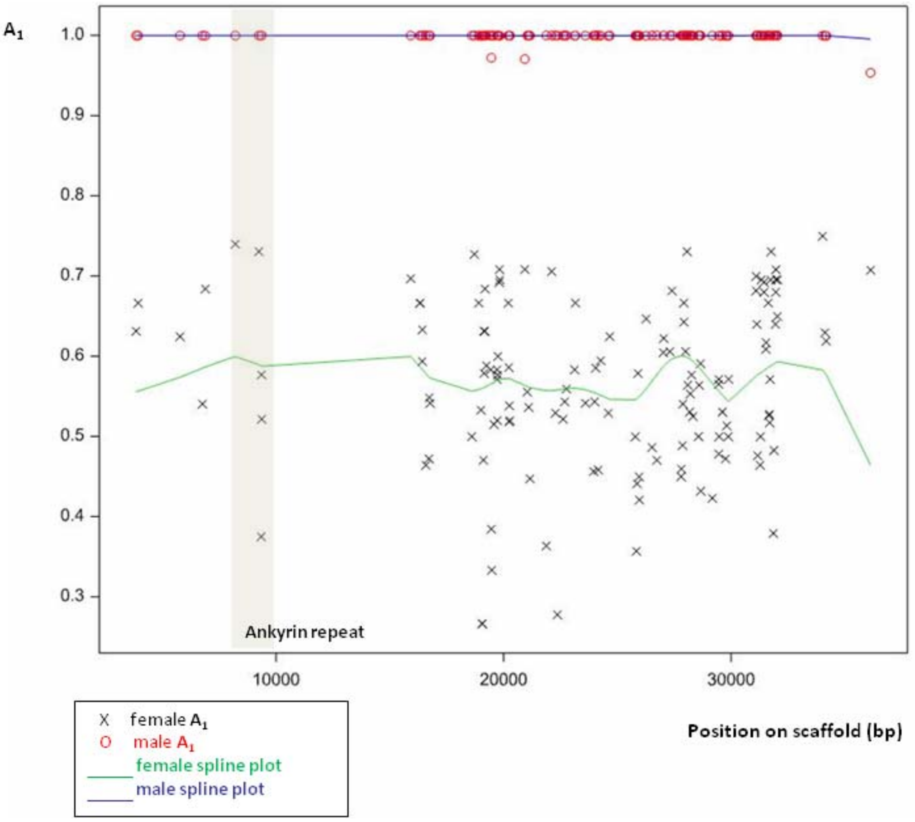


Figure 3-10. Scaffold_2087, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Genes are shown in grey: a gene containing Ankyrin repeat (artfr2087g00020).

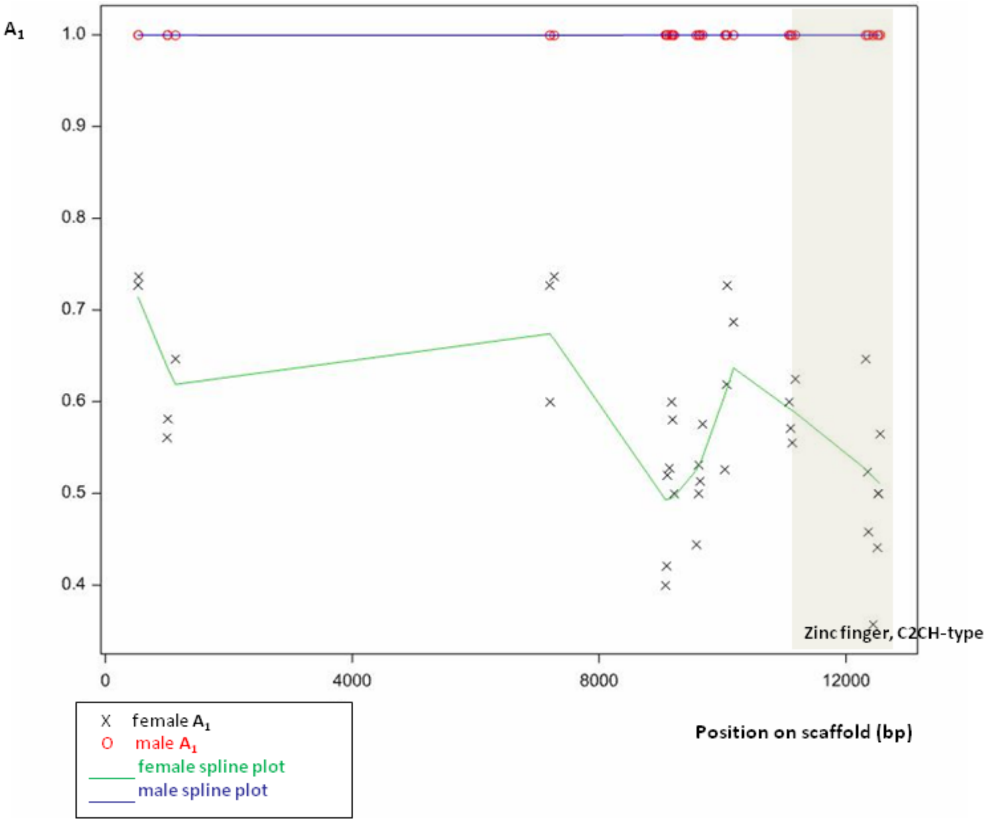


Figure 3-11. Scaffold_21607, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Genes are shown in grey: *Zinc finger, C2CH-type* (artfr21607g00030).

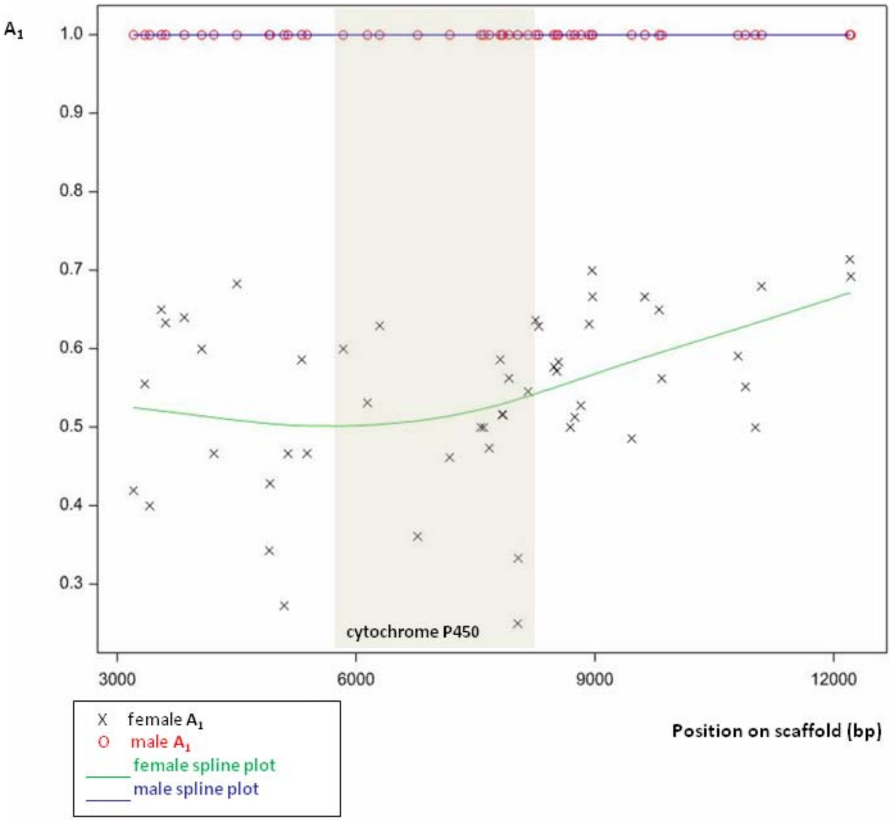


Figure 3-12. Scaffold_23412, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Genes are shown in grey: *Cytochrome P450* (artfr23412g00030).

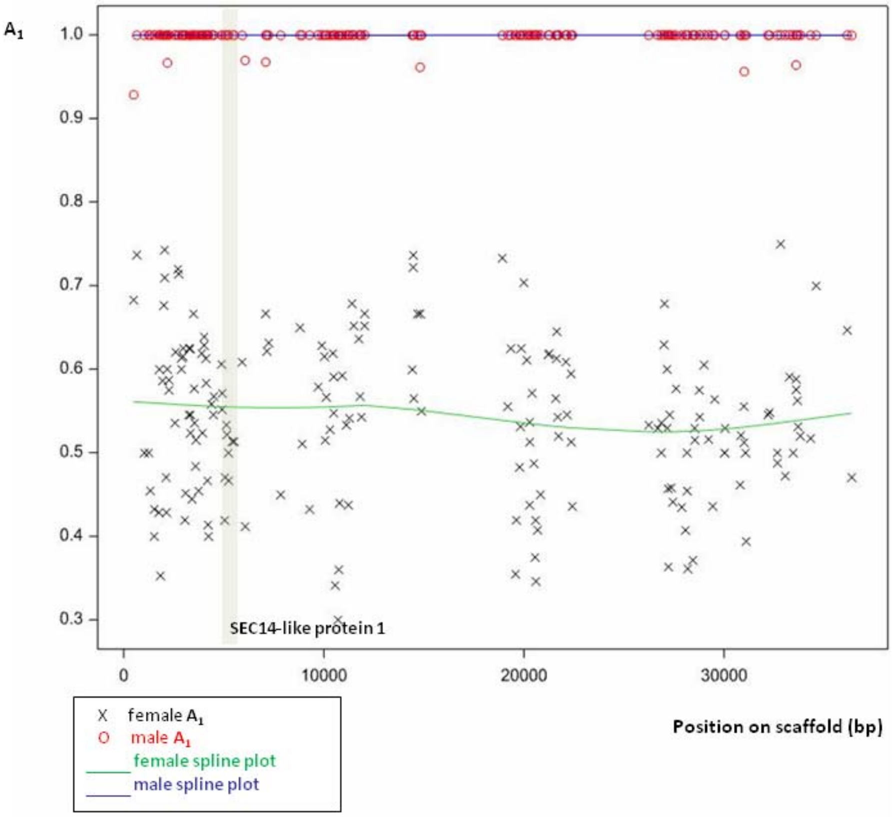
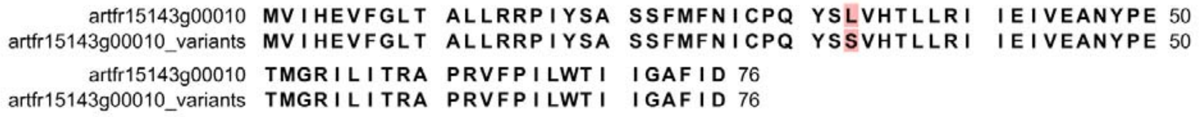


Figure 3-13. Scaffold_4146, reference allele frequency A_1 of selected SNPs in female and male reads with respective spline plots. Gene *SEC-14-like protein 1* (artfr4146g00010) is shown in grey.

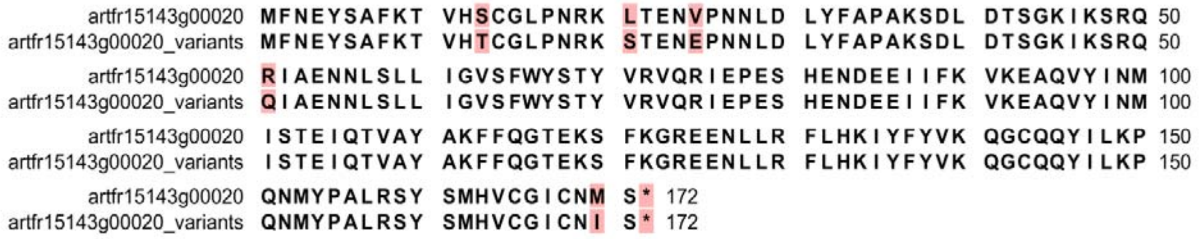
3.4.3. Sex-specific genes with non-synonymous SNPs

In the eight genes (Table 3-5, blue font) with only detected SNPs, heterozygous in females and homozygous in males, synonymous and non-synonymous selected SNPs were identified (Table 3-6) by alignment of the reference protein and its translated variant (Figures 3-14 to 3-19).



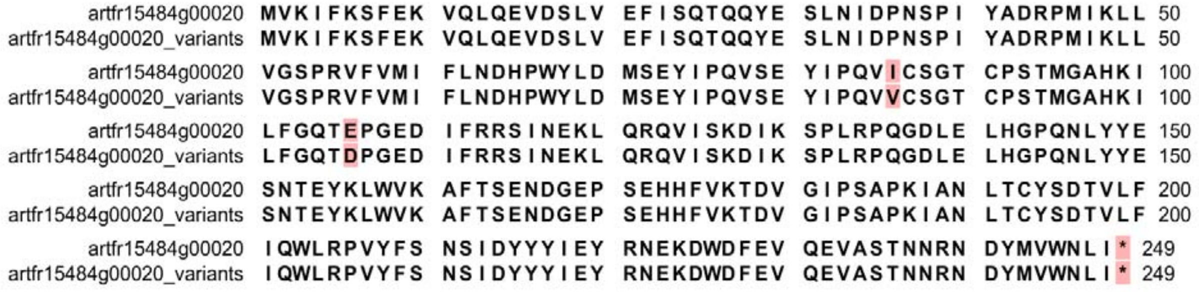
```
         artfr15143g00010 MVIHEVFGLT ALLRRPIYSA SSFMFNICPQ YSLVHTLLRI IEIVEANYPE 50
artfr15143g00010_variants MVIHEVFGLT ALLRRPIYSA SSFMFNICPQ YSSVHTLLRI IEIVEANYPE 50
         artfr15143g00010 TMGRILITRA PRVFPILWTI IGAFID 76
artfr15143g00010_variants TMGRILITRA PRVFPILWTI IGAFID 76
```

Figure 3-14. Alignment of *A. franciscana* reference protein artfr15143g00010 (from a gene containing the CRAL-TRIO domain) and its variant protein (artfr15143g00010_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.



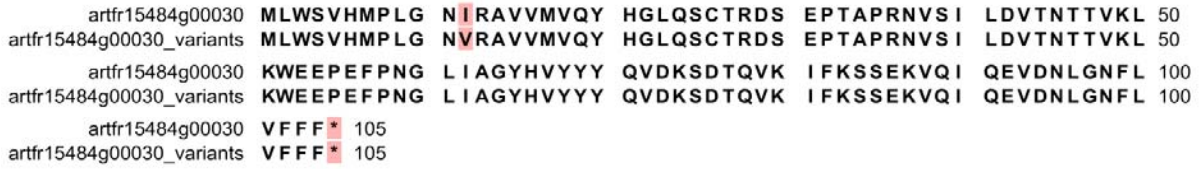
```
         artfr15143g00020 MFNEYSAFKT VHSCGLPNRK LTENVPNNLD LYFAPAKSDL DTSGKIKSRQ 50
artfr15143g00020_variants MFNEYSAFKT VHTCGLPNRK STENEPNNLD LYFAPAKSDL DTSGKIKSRQ 50
         artfr15143g00020 RIAENNLSLL IGVSFWYSTY VRVQRIEPES HENDEEIIFK VKEAQVYINM 100
artfr15143g00020_variants QIAENNLSLL IGVSFWYSTY VRVQRIEPES HENDEEIIFK VKEAQVYINM 100
         artfr15143g00020 ISTEIQTVAY AKFFQGTEKS FKGREENLLR FLHKIYFYVK QGCQQYILKP 150
artfr15143g00020_variants ISTEIQTVAY AKFFQGTEKS FKGREENLLR FLHKIYFYVK QGCQQYILKP 150
         artfr15143g00020 QNMYPALRSY SMHVCGICNM S* 172
artfr15143g00020_variants QNMYPALRSY SMHVCGICNI S* 172
```

Figure 3-15. Alignment of *A. franciscana* reference protein artfr15143g00020 (from *F0F1 ATP synthase subunit beta*) and its variant protein (artfr15143g00020_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.



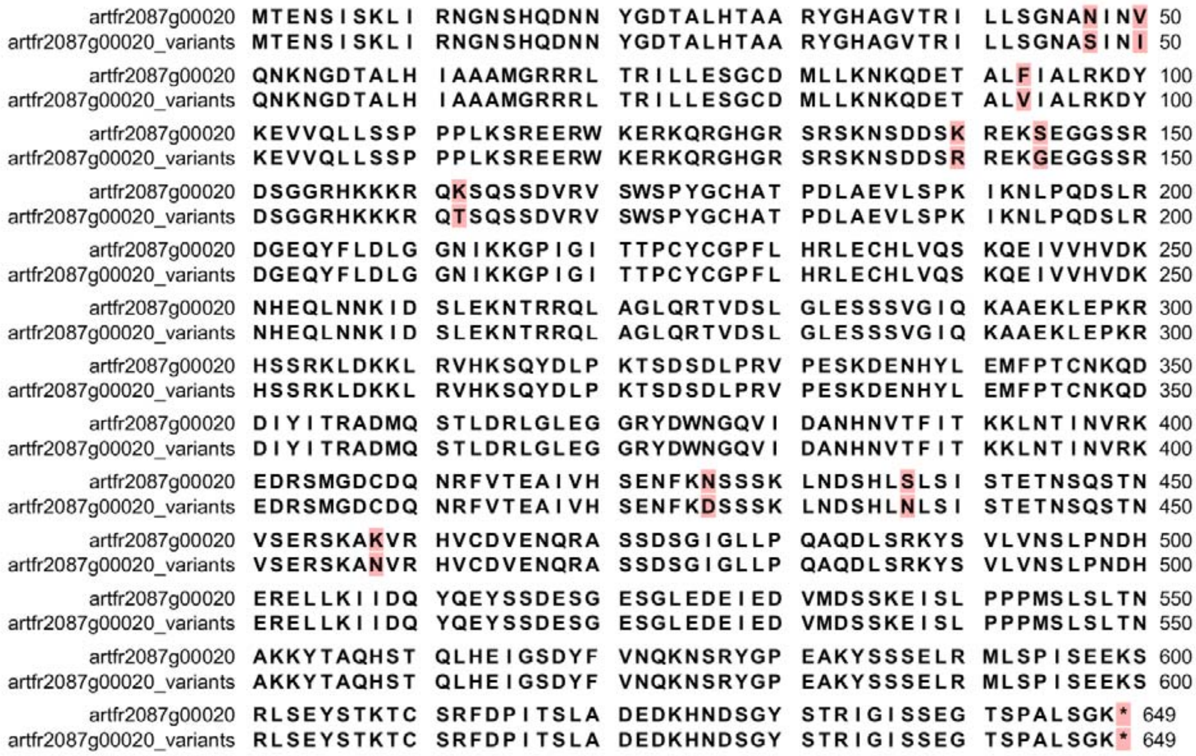
```
         artfr15484g00020 MVKIFKSFEK VQLQEVDSLV EFISQTQQYE SLNIDPNSPI YADRPMIKLL 50
artfr15484g00020_variants MVKIFKSFEK VQLQEVDSLV EFISQTQQYE SLNIDPNSPI YADRPMIKLL 50
         artfr15484g00020 VGSPRVFVMI FLNDHPWYLD MSEYIPQVSE YIPQVICSGT CPSTMGAHKI 100
artfr15484g00020_variants VGSPRVFVMI FLNDHPWYLD MSEYIPQVSE YIPQVVCSGT CPSTMGAHKI 100
         artfr15484g00020 LFGQTEPGED IFRRSINEKL QRQVISKDIK SPLRPQGDLE LHGPQNLYYE 150
artfr15484g00020_variants LFGQTDPGED IFRRSINEKL QRQVISKDIK SPLRPQGDLE LHGPQNLYYE 150
         artfr15484g00020 SNTEYKLWVK AFTSENDGEP SEHHFVKTDV GIPSAPKIAN LTCYSDTVLF 200
artfr15484g00020_variants SNTEYKLWVK AFTSENDGEP SEHHFVKTDV GIPSAPKIAN LTCYSDTVLF 200
         artfr15484g00020 IQWLRPVYFS NSIDYYYIEY RNEKDWDFEV QEVASTNNRN DYMVWNLI* 249
artfr15484g00020_variants IQWLRPVYFS NSIDYYYIEY RNEKDWDFEV QEVASTNNRN DYMVWNLI* 249
```

Figure 3-16. Alignment of *A. franciscana* reference protein artfr15484g00020 (from *Fibronectin, type III*) and its variant protein (artfr15484g00020_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.



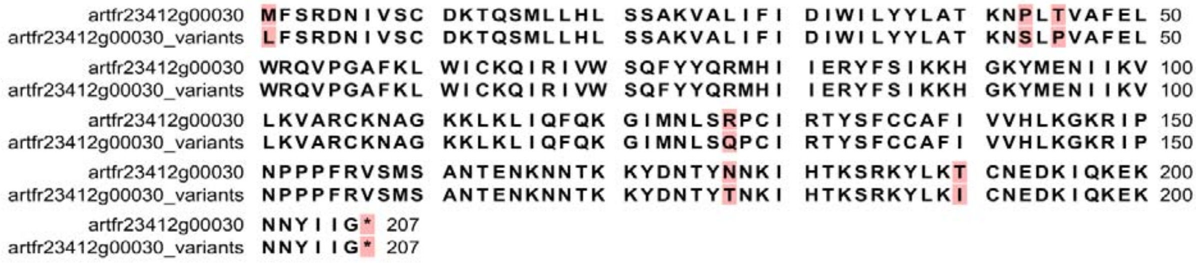
```
         artfr15484g00030 MLWSVHMPLG NIRAVVMVQY HGLQSCTRDS EPTAPRNVSI LDVTNTTVKL 50
artfr15484g00030_variants MLWSVHMPLG NVRAVVMVQY HGLQSCTRDS EPTAPRNVSI LDVTNTTVKL 50
         artfr15484g00030 KWEEPEFPNG LIAGYHVYYY QVDKSDTQVK IFKSSEKVQI QEVDNLGNFL 100
artfr15484g00030_variants KWEEPEFPNG LIAGYHVYYY QVDKSDTQVK IFKSSEKVQI QEVDNLGNFL 100
         artfr15484g00030 VFFF* 105
artfr15484g00030_variants VFFF* 105
```

Figure 3-17. Alignment of *A. franciscana* reference protein artfr15484g00030 (from *Fibronectin, type III*) and its variant protein (artfr15484g00030_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.



```
artfr2087g00020           MTENSISKLI RNGNSHQDNN YGDTALHTAA RYGHAGVTRI LLSGNANINV 50
artfr2087g00020_variants  MTENSISKLI RNGNSHQDNN YGDTALHTAA RYGHAGVTRI LLSGNASINI 50
artfr2087g00020           QNKNGDTALH IAAAMGRRRL TRILLESGCD MLLKNKQDET ALFIALRKDY 100
artfr2087g00020_variants  QNKNGDTALH IAAAMGRRRL TRILLESGCD MLLKNKQDET ALVIALRKDY 100
artfr2087g00020           KEVVQLLSSP PPLKSREERW KERKQRGHGR SRSKNSDDSK REKSEGGSSR 150
artfr2087g00020_variants  KEVVQLLSSP PPLKSREERW KERKQRGHGR SRSKNSDDSR REKGEGGSSR 150
artfr2087g00020           DSGGRHKKKR QKSQSSDVRV SWSPYGCHAT PDLAEVLSPK IKNLPQDSLR 200
artfr2087g00020_variants  DSGGRHKKKR QTSQSSDVRV SWSPYGCHAT PDLAEVLSPK IKNLPQDSLR 200
artfr2087g00020           DGEQYFLDLG GNIKKGPIGI TTPCYCGPFL HRLECHLVQS KQEIVVHVDK 250
artfr2087g00020_variants  DGEQYFLDLG GNIKKGPIGI TTPCYCGPFL HRLECHLVQS KQEIVVHVDK 250
artfr2087g00020           NHEQLNNKID SLEKNTRRQL AGLQRTVDSL GLESSSVGIQ KAAEKLEPKR 300
artfr2087g00020_variants  NHEQLNNKID SLEKNTRRQL AGLQRTVDSL GLESSSVGIQ KAAEKLEPKR 300
artfr2087g00020           HSSRKLDKKL RVHKSQYDLP KTSDSDLPRV PESKDENHYL EMFPTCNKQD 350
artfr2087g00020_variants  HSSRKLDKKL RVHKSQYDLP KTSDSDLPRV PESKDENHYL EMFPTCNKQD 350
artfr2087g00020           DIYITRADMQ STLDRLGLEG GRYDWNGQVI DANHNVTFIT KKLNTINVRK 400
artfr2087g00020_variants  DIYITRADMQ STLDRLGLEG GRYDWNGQVI DANHNVTFIT KKLNTINVRK 400
artfr2087g00020           EDRSMGDCDQ NRFVTEAIVH SENFKNSSSK LNDSHLSLSI STETNSQSTN 450
artfr2087g00020_variants  EDRSMGDCDQ NRFVTEAIVH SENFKDSSSK LNDSHLNLSI STETNSQSTN 450
artfr2087g00020           VSERSKAKVR HVCDVENQRA SSDSGIGLLP QAQDLSRKYS VLVNSLPNDH 500
artfr2087g00020_variants  VSERSKANVR HVCDVENQRA SSDSGIGLLP QAQDLSRKYS VLVNSLPNDH 500
artfr2087g00020           ERELLKIIDQ YQEYSSDESG ESGLEDEIED VMDSSKEISL PPPMSLSLTN 550
artfr2087g00020_variants  ERELLKIIDQ YQEYSSDESG ESGLEDEIED VMDSSKEISL PPPMSLSLTN 550
artfr2087g00020           AKKYTAQHST QLHEIGSDYF VNQKNSRYGP EAKYSSSELR MLSPISEEKS 600
artfr2087g00020_variants  AKKYTAQHST QLHEIGSDYF VNQKNSRYGP EAKYSSSELR MLSPISEEKS 600
artfr2087g00020           RLSEYSTKTC SRFDPITSLA DEDKHNDSGY STRIGISSEG TSPALSGK* 649
artfr2087g00020_variants  RLSEYSTKTC SRFDPITSLA DEDKHNDSGY STRIGISSEG TSPALSGK* 649
```

Figure 3-18. Alignment of *A. franciscana* reference protein artfr2087g00020 (from a gene containing Ankyrin repeat) and its variant protein (artfr2087g00020_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.



```
artfr23412g00030           MFSRDNIVSC DKTQSMLLHL SSAKVALIFI DIWILYYLAT KNPLTVAFEL 50
artfr23412g00030_variants  LFSRDNIVSC DKTQSMLLHL SSAKVALIFI DIWILYYLAT KNSLPVAFEL 50
artfr23412g00030           WRQVPGAFKL WICKQIRIVW SQFYYQRMHI IERYFSIKKH GKYMENIIKV 100
artfr23412g00030_variants  WRQVPGAFKL WICKQIRIVW SQFYYQRMHI IERYFSIKKH GKYMENIIKV 100
artfr23412g00030           LKVARCKNAG KKLKLIQFQK GIMNLSRPCI RTYSFCCAFI VVHLKGKRIP 150
artfr23412g00030_variants  LKVARCKNAG KKLKLIQFQK GIMNLSQPCI RTYSFCCAFI VVHLKGKRIP 150
artfr23412g00030           NPPPFRVSMS ANTENKNNTK KYDNTYNNKI HTKSRKYLKT CNEDKIQKEK 200
artfr23412g00030_variants  NPPPFRVSMS ANTENKNNTK KYDNTYTNKI HTKSRKYLKI CNEDKIQKEK 200
artfr23412g00030           NNYIIG* 207
artfr23412g00030_variants  NNYIIG* 207
```

Figure 3-19. Alignment of *A. franciscana* reference protein artfr23412g00030 (from *Cytochrome P450*) and its variant protein (artfr23412g00030_variants) containing all selected SNPs. Variant amino acids are highlighted in orange.

Two genes showed no SNPs in the exonic region (*zinc finger C2H2 type* and *SEC14-like protein 1*). The other six genes (CRAL-TRIO domain; *F0F1 ATP synthase subunit beta*; *Fibronectin type III*; Ankyrin repeat and *Cytochrome P450*) each contained at least one non-synonymous SNP, with *Cytochrome P450* containing only non-synonymous SNPs, of which one in the start codon, resulting in a shorter protein in males.

Table 3-6. Synonymous and non-synonymous selected SNPs in the eight genes showing only detected SNPs, heterozygous in females and homozygous in males. The gene ID from the Orcae online annotation platform is provided. It consists of: "artfr" for *Artemia franciscana*, the scaffold number and "g", followed by the gene number. A functional description of the gene, the position of the SNP in the scaffold, its reference (Ref) and variant (Var) nucleotides (nt) and synonymous (SYN) or non-synonymous (NON) status of the SNP and the (Ref) and variant (Var) amino acids (aa) in the resulting protein and changes in secondary protein structure are indicated as well.

Gene ID	Functional description	Position	Ref nt	Var nt	SYN/NON	Ref aa	Var aa	Protein secondary structure
artfr15143g00010	*CRAL-TRIO domain*	6174	A	G	NON	L	S	change
		6419	G	C	SYN	_	_	_
artfr15143g00020	*F0F1 ATP synthase subunit beta*	8739	C	T	NON	S	T	change
		9490	A	G	SYN	_	_	_
		9509	C	T	NON	L	S	_
		9627	A	C	SYN	_	_	_
		9628	A	T	NON	V	E	_
		9640	A	G	NON	R	Q	_
		9664	C	G	NON	M	I	change
artfr15484g00020	*Fibronectin, type III*	7842	C	T	SYN	_	_	_
		7989	G	C	SYN	_	_	_
		8697	T	A	SYN	_	_	_
		10062	T	G	NON	I	V	_
		10222	T	C	NON	E	D	_
artfr15484g00030	*Fibronectin, type III*	11187	C	G	SYN	_	_	_
		11708	T	C	NON	I	V	_
artfr2087g00020	Ankyrin repeat	9351	T	C	SYN	_	_	_
		9362	A	G	NON	N	S	_
		9370	G	A	NON	V	I	_
		16581	T	G	NON	F	V	_
		16723	A	G	NON	K	R	change
		16725	C	A	SYN	_	_	_
		16734	A	G	NON	S	G	_
		16769	A	G	SYN	_	_	_
		20282	A	C	NON	K	T	_
		28178	A	G	SYN	_	_	_
		28202	A	G	SYN	_	_	_
		28259	T	C	SYN	_	_	_
		28322	C	G	SYN	_	_	_
		28566	A	G	NON	N	D	_
		28571	C	T	SYN	_	_	_
		28600	G	A	NON	S	N	_
		28655	G	C	SYN	_	_	_
		28664	A	T	NON	K	N	change
		29174	A	T	SYN	_	_	_
artfr23412g00030	*cytochrome P450*	6143	G	A	NON	M	L	change
		6296	T	G	NON	P	S	change
		7176	C	T	NON	T	P	_
		8030	T	G	NON	R	Q	_
		8036	G	A	NON	N	T	_
		8162	T	A	NON	T	I	_
artfr21607g00030	*Zinc finger, C2CH-type*	No SNPs in exonic region						
artfr4146g00010	*SEC14-like protein 1*	No SNPs in exonic region						

Changes in secondary protein structure, such as disappearing elements or alpha-helix elements turning into beta-sheet elements due to non-synonymous SNPs were found in *CRAL-TRIO domain*; *F0F1 ATP synthase subunit beta*; Ankyrin repeat and *Cytochrome P450*, (Figures 3-20 to 3-23) but not in *Fibronectin type III* (Table 3-3).

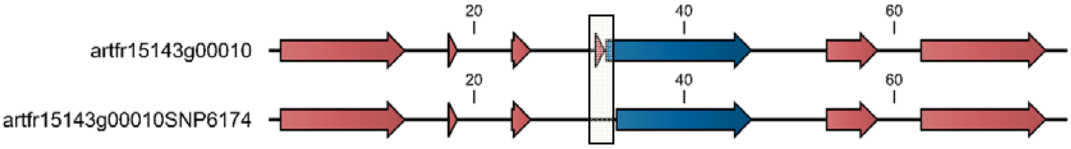


Figure 3-20. Changes in secondary protein structure (grey box) due to non-synonymous SNPs in *CRAL-TRIO domain* (artfr15143g00010). Top: the reference protein, below, the variant protein. The variant protein is named after the reference protein, followed by "SNP" and the SNP position in the *Artemia* genome scaffold. Length is indicated in amino acids. Predicted elements are alpha-helix (blue) and beta-sheet (red).

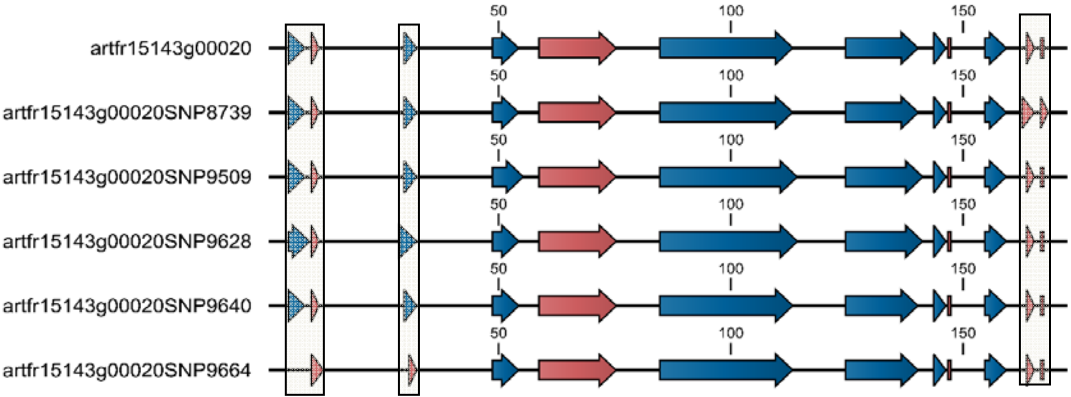


Figure 3-21. Changes in secondary protein structure (grey boxes) due to non-synonymous SNPs in *F0F1 ATP synthase subunit beta* (artfr15143g00020). Top: the reference protein, below, the variant proteins. Each variant protein is named after the reference protein, followed by "SNP" and the SNP position in the *Artemia* genome scaffold. Length is indicated in amino acids. Predicted elements are alpha-helix (blue) and beta-sheet (red).

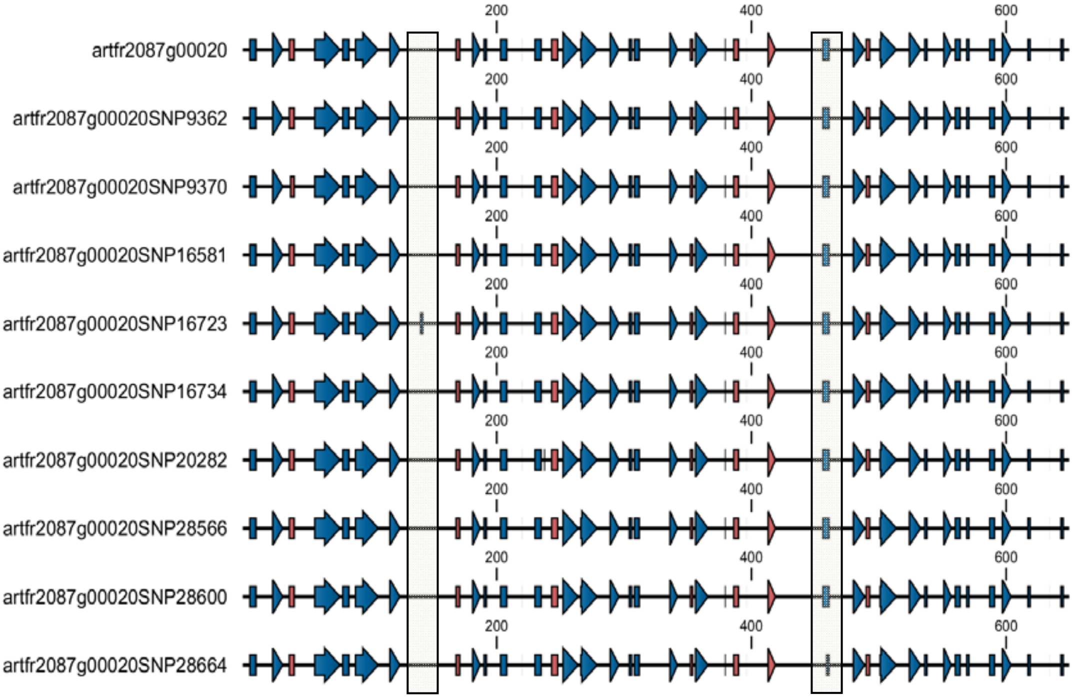


Figure 3-22. Changes in secondary protein structure (grey boxes) due to non-synonymous SNPs in Ankyrin repeat (artfr2087g00020). Top: the reference protein, below, the variant proteins. Each variant protein is named after the reference protein, followed by "SNP" and the SNP position in the *Artemia* genome scaffold. Length is indicated in amino acids. Predicted elements are alpha-helix (blue) and beta-sheet (red).

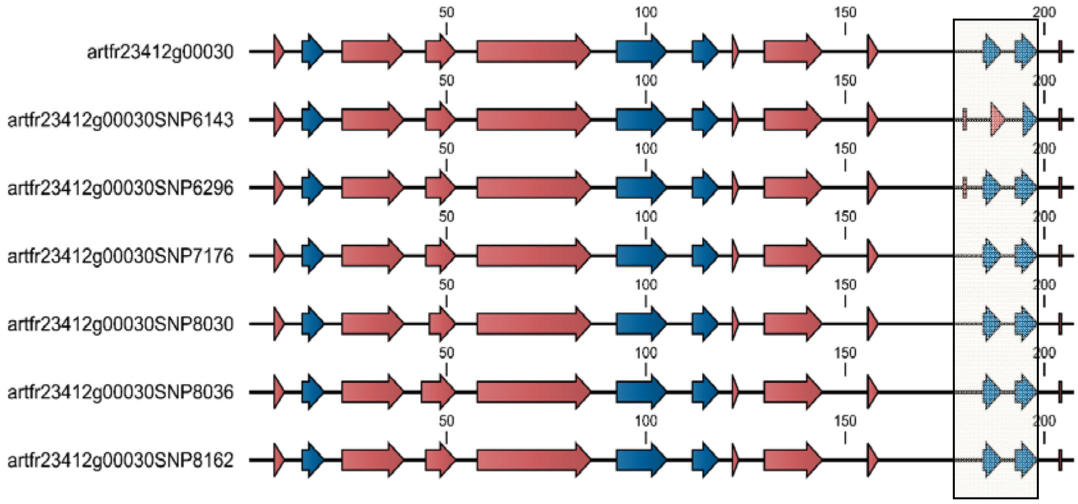


Figure 3-23. Changes in secondary protein structure (grey box) due to non-synonymous SNPs in *Cytochrome P450* (artfr23412g00030). Top: the reference protein, below, the variant proteins. Each variant protein is named after the reference protein, followed by "SNP" and the SNP position in the *Artemia* genome scaffold. Length indications are expressed in amino acids. Predicted elements are alpha-helix (blue) and beta-sheet (red).

3.4.4. Arthropod sex-determining gene homologs, not selected by BSA

Genes, homologous to known insect sex-determining genes such as *dsx*, *SRY*, *sex-lethal*, *fem-1*, *tra-1*, *tra-2*, *fruitless*, *runt*, *deadpan*, *daughterless*, *extra macrochaetae*, *groucho* and *sans fille* (section 1.3.) were found in the *Artemia* genome (Table 3-7, Figure 3-24). None of these genes were selected with the BSA strategy.

Table 3-7 Genes homologous to known insect sex-determining genes found in the *Artemia* genome. These genes were not selected with the BSA strategy. Gene characteristics are provided from the Orcae online annotation platform. The gene ID consists of: "artfr", which stands for *Artemia franciscana*, the scaffold number and "g", followed by the gene number. The start and end positions of the gene on the scaffold, the functional description of the gene and the homologous protein from the best BLAST (blastx) hit of the NCBI "nr database" with the gene are provided as well, completed by ID, E-value and organism of the BLAST hit.

Gene ID	Functional description	Homologous protein	ID	E-value	Organism
artfr14066g00010	DM DNA-binding domain	*double sex and Mab-3 related transcription factor 2 (dsx)*	0.68	1.00E-29	*Pinctada martensi*
artfr7572g00040	unnamed product	*SRY-related HMG box C protein (SRY)*	0.51	6.00E-45	*Platynereis dumerilii*
artfr13354g00020	Ankyrin repeat	*fem-1-like protein*	0.45	3.00E-96	*Daphnia pulex*
artfr3457g00010	Ankyrin repeat	*sex-determining protein fem-1*	0.65	5.00E-40	*Scylla paramamosain*
artfr14227g00040	Nucleotide-binding, alpha-beta plait	*transformer-2 sex-determining protein, putative (tra-2)*	0.67	2.00E-29	*Pediculus humanus corporis*
artfr9384g00030	unnamed product	*transformer-1-like protein (tra-1)*	0.88	1.00E-20	*Daphnia pulex*
artfr1775g00110	hypothetical protein EAI_05926	*transformer (tra)*	0.41	6.00E-06	*Daphnia magna*
artfr2959g00130	unnamed product	*fruitless (fru)*	0.51	1.00E-33	*Drosophila virilis*
artfr41457g00010	unnamed product	*transcription factor runt-like protein (runt)*	0.63	5.00E-48	*Daphnia pulex*
artfr155g00090	Myc-type, basic helix-loop-helix (bHLH) domain	*Protein deadpan (dpn)*	0.53	3.00E-30	*Acromyrmex echinatior*
artfr14134g00020	Myc-type, bHLH domain	*daughterless-like protein, isoform 2 (da)*	0.45	6.00E-44	*Daphnia pulex*
artfr15575g00030	Myc-type, bHLH domain	*extra macrochaetae-like protein (emc)*	0.60	1.70E-02	*Daphnia pulex*
artfr2769g00040	unnamed product	*groucho-like protein (gro)*	0.86	1.00E-51	*Daphnia pulex*
artfr2996g00020	unnamed product	*sans fille-like protein (snf)*	0.85	1.00E-23	*Daphnia pulex*
artfr411g00050	sex-lethal	-	-	-	-

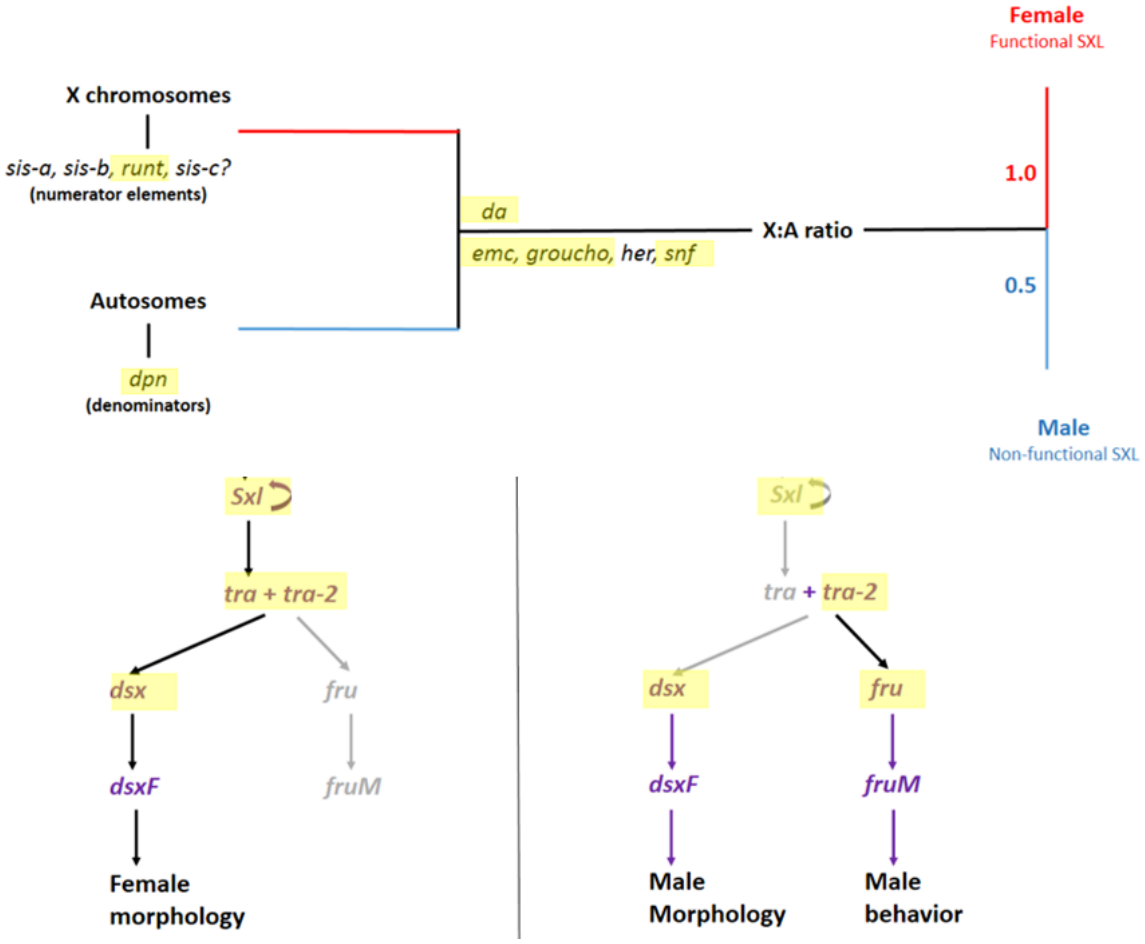


Figure 3-24. Sex determination in *Drosophila* from the chromosomal signaling cascade until *sxl* (top figure [132].) and from *sxl* until *dsx* (bottom figure [106]). Genes showing homology with the *Artemia* genome in the annotation platform Orcae are shown in yellow on the sex determination pathways of *Drosophila*. Only genes *sisterless* (*sis-a*, *sis-b* and *sis-c*) and *hermaphrodite* (*her*) did not show any homology within the *Artemia* genome.

Homology within the *Artemia* genome was also found for some of the crustacean sex-related genes reviewed in 3.2.2.2.: *fushi tarazu*, *forkhead box*, *WNT*, *argonaute*, *testis-specific*, *VASA*, *SOX*, *star*, *ECM*, *tudor*, *GATA*, *prohibitin* and *Cytochrome P450*. Of these genes, only *Cytochrome P450* was selected with the BSA strategy.

3.5. Discussion

Here, BSA revealed eight candidate primary sex-determining genes in the *Artemia* genome, including four genes with exonic non-synonymous SNPs that affected secondary protein structure: *Cytochrome P450*, *F0F1 ATP synthase subunit beta* and two genes respectively containing a CRAL-TRIO domain and ankyrin repeat, two *Fibronectin* genes with exonic non-synonymous SNPs that had no effect on secondary protein structure and finally *SEC14* and *Zinc finger C2H2-type* that contained no exonic SNPs.

Several genes, not selected by BSA, but highly homologous with insect sex-determining genes (*doublesex*, *sex-determining region Y*, *sex-lethal*, *feminization 1*, *transformer 1*, *transformer 2*, *fruitless*, *runt*, *deadpan*, *daughterless*, *extra macrochaetae*, *groucho* and *sans fille*) or with crustacean sex-related genes (*extra macrochaetae*, *fushi tarazu*, *forkhead box*, *WNT*, *argonaute*, *testis-specific*, *VASA*, *SOX*, *star*, *ECM*, *tudor*, *GATA*, *prohibitin* and *Cytochrome P450*) were present in the *Artemia* genome as well.

3.5.1. AFLP analysis, sequencing and BLAST of sex-linked markers

Scaffold_13, containing the sex-linked AFLP markers, was selected with the BSA strategy, revealing that the BSA selection strategy picked up scaffolds carrying shown to be sex-linked by AFLP. Although scaffold_13 did not contain selected SNPs within the CDS of its genes, it did contain a selected SNP and two indels in the area to which the sex-linked AFLP marker E112M223M273.0F mapped (Figure 3-6). The absence of SNPs on the remaining part of scaffold_13 and in the sex-linked markers (Table 3-2), except the above mentioned marker, could be explained partly by the occurrence of single-sex coverage (Figure 3-7), because to detect a SNP with VarScan, it must differ in males and females, thus requiring coverage from both sexes.

3.5.2. The sex-determining system of *Artemia*

The hypothesis of a primary sex-determining gene lying on both the Z and the W chromosomes (Figure 3-3) is the most likely system, because it concurs with both the results from the *A. franciscana* AFLP linkage maps and from the BSA: sex-linked markers with the proposed segregation were found (Table 2-3, pattern 1), as well as a SNP and two indels, heterozygous in females and homozygous in males, present on sex-linked AFLP marker E112M223M273.0F (Table 2-3, pattern 1), covered by male and female reads. In either case, the sex-linked AFLP markers, as well as the SNPs identified with BSA lead to a region close to the sex-determining gene.

3.5.3. Bulked segregant analysis by whole-genome sequencing

In the spline plots of the eight identified genes that contain mostly or all SNPs heterozygous in females and homozygous in males (Table 3-5, blue, Figures 3-8 to 3-13), the reference allele frequency A_1 in females is mostly over 0.5, possibly due to stringent BWA mapping, slightly favoring the reference over the variant allele. Default mismatches per read have been allowed through the parameter settings of BWA. This situation did not influence the overall results.

SEC14 and *Fibronectin III* might be duplicated genes or be present more than once in the assembly, due to fragmentation in the *Artemia* assembly (see Chapter 4). *SEC14* is represented twice (Table 3-5): one candidate gene has its functional description, whereas another is highly homologous with *SEC14-like protein 1*. *Fibronectin III* is represented twice among the selected candidate genes as well.

The twelve BSA-identified genes with functional description linked to sex determination or differentiation in the literature (Table 3-5) are discussed. Of these genes, the following four candidate primary sex-determining genes had non-synonymous SNPs, causing production of sex-specific proteins with a different secondary structure in males and females.

- The *Cytochrome P450* gene superfamily (*CYPs*) (Table 3-5) has a diverse range of functions in animals. Some *CYPs* are regulated in a sexually dimorphic fashion through endocrine mechanisms. A subgroup of P450 steroid hydroxylases known as aromatases are expressed in regions of the gonads and brain, important for the neuroendocrine regulation of reproduction, fertility and sexual behavior [15]. Epigenetic mechanisms involving DNA methylation, histone deacetylation and histone methylation of *Cytochrome P450 aromatase* (encoded by *cyp19a1*) have been suggested to drive natural sex changes in teleosts and gonadal differentiation in other vertebrates [235]. *Cytochrome P450 aromatase* has also been identified in *Macrobrachium nipponense* as a putative gene related to sex determination/differentiation through transcriptome analysis and pathway mapping [124].

- The CRAL-TRIO domain containing gene is highly homologous to *SEC 14* (Table 3-5). Exposure of *D. magna* to JH agonists (methylfarnesoate, JH III, methoprene, and the insecticides pyriproxyfen and fenoxycarb [199]), reduces the reproductive function, causes production of male offspring and downregulates *SEC14* expression.

- Ankyrin repeat (Table 3-5) is a repeat, present in the putative *fem-1* gene in the monarch butterfly *Danaus plexippus* and in *C. elegans fem-1* and *fem-3*. It occurs also in the *Drosophila Notch* gene, responsible for cell fate determination in early development, specifically in embryogenesis [222]. Ankyrin is, however a very common repeat, present in many proteins that fulfill several functions in different organisms.

- *F0F1 ATP synthase subunit beta* (Table 3-5) is primarily known as a metabolic gene, but has been shown to be differentially expressed in male and female mussel eggs (*Mytilus edulis*) [49].

Of the 12 BSA-identified genes, the two following candidate primary sex-determining genes had non-synonymous SNPs, causing production of sex-specific proteins, but with the same secondary structure in males and females.

-In *Artemia,* two genes with a Fibronectin III domain (Table 3-5) were detected. Fibronectin domains are found in a wide variety of extracellular animal proteins, such as *kal-1* in *Drosophila* [11]. In *Drosophila*, *kal-1* is expressed during morphogenetic processes in the embryonic development and sex-specifically in male-specific somatic gonadal precursor cells [11]. Recently, a Fibronectin III domain-containing gene was found to have ovary-enhanced expression in dragon fish (*Scleropages formosus*) [184]. One of the Fibronectin-containing genes found in *Artemia* is homologous with *S-adenosyl-methyltransferase* (*mraw*) that is involved in germplasm formation in *Drosophila* [28].

Of the 12 BSA-identified genes, the following two candidate primary sex-determining genes had no SNPs in the exonic region, but they had an interesting connection with sex determination in literature. It could be that sex-specific regions are present in the gene regulatory regions, which were not analyzed within the scope of this work.

- In *Bombyx mori*, the best-studied sex determination model for arthropod female heterogametes (WZ/ZZ), *Zinc finger C2H2-type* (Table 3-5) and *C3H*-type genes are found in tandem repeats within the minimal *Fem* region on chromosome W and are expressed during early embryonic development, prior to *B. mori dsx* expression [146]. Therefore, *C2H2* and *C3H-type* zinc finger proteins are currently the number one candidate genes for the source of the initial signal for female development in *B. mori*.

- Exposure of *D. magna* to JH agonists (methylfarnesoate, JH III, methoprene, and the insecticides pyriproxyfen and fenoxycarb [199]), reduces the reproductive function, causes production of male offspring and downregulates *SEC14 (*Table 3-5) expression.

Of the 12 BSA-identified genes, contrary to the eight previously discussed candidate primary sex-determining genes, the following four genes did not show all SNPs heterozygous in females and homozygous in males (Table 3-5, black), but they had a connection with sex determination in literature.

- In *C. elegans*, *sel-10* encodes a WD40-repeat-containing F-box protein (Table 3-5) that probably mediates the degradation of important sex determination factors (*fem*) [87,103]. In expression studies of ovary tissue in *M. rosenbergii*, WD repeat-containing protein is among the most abundant transcripts [92].

- *Him17* is a Zinc finger, C2CH-type protein (Table 3-5) responsible for meiotic silencing in nematodes, because H3K9 methylation of the male X is defective in mutants [100].

- Histone core (Table 3-5) proteins code for core histones H2A, H2B, H3 and H4. Activity of *C. elegans* X chromosomes at the germline stages is caused by different methylation patterns of core histone H3 [170]. Establishment of germline sexual identity is critical for production of male and female germline stem cells, and of sperm vs. eggs.

- The gene containing a protein kinase-like domain (Table 3-5) is highly homologous to *DYRK* (*dual-specificity tyrosine-phosphorylation-regulated kinase*). This gene has been found to regulate pre-mRNA splicing in spermatogonia and proliferation of spermatogonia and Sertoli cells in mice [54].

The identified candidate genes often present domains of or homology to genes that are linked with sex determination. Seemingly, there could be pathway connections between the selected genes, because in Table 3-5, two candidate genes (Fibronectin III and the gene containing a protein kinase-like domain) are highly homologous with *tyrosine-protein phosphatase 99A* and *dual specificity tyrosine-phosphorylation-regulated kinase* (DYRK), respectively. Two zinc finger proteins are present in the listed candidate sex-determining genes as well (*Zinc finger C2H2-type* and *Zinc finger, C2CH-type*).

Epigenetic regulatory mechanisms are known to contribute to sex determination and reproductive organ formation in plants, invertebrates, and vertebrates [160]. The three main epigenetic mechanisms for gene expression regulation include DNA methylation, histone modifications, and non-coding RNAs. Among the candidate genes for sex determination in *Artemia*, there is a high incidence of genes involved in epigenetic processes, such as methylation and histone modifications (*Zinc finger, C2CH-type*; *Cytochrome P450 aromatase*; *histone core* and *Fibronectin III*).

It should be considered that this work focused on genes with functional descriptions involved in sex determination in other organisms. In other words, interesting or important genes might still be found among the genes lacking a description, but with a known domain or homologous protein. Also, selection of genes based on sex-determining functionality in other organisms could not cover all genes of interest, because gene functions may vary among genera and conservation of primary sex-determining genes is generally low. The position of a homologous gene in the sex determination cascade may differ among closely related genera as well, as seen in *Apis*, where *Fem* is not a primary sex-determining gene, in contrast to other insects. Also, analysis of indels in genes and gene regulatory regions and of SNPs in regulatory regions could lead to the discovery of more genes that cause production of sex-specific proteins. Finally, manual curation of all the genes thought to be involved in sex determination could improve the functional annotation and possibly the recognition of sex-determining genes.

3.5.4. Arthropod sex-determining gene homologs, not selected by BSA

Genes, homologous to known insect sex-determining genes (*dsx*, *fru*, *SRY*, *sxl*, *tra-1*, *tra-2*, *fem-1*, *runt*, *dpn*, *da*, *emc*, *gro* and *sf*), representing most genes in the *Drosophila* sex determination cascade (Figure 3-24) are present in the *Artemia* genome, suggesting that *Artemia* might use similar pathways. These genes were, however, not selected by BSA and have thus far not been functionally validated in *Artemia.* It could be that these genes are not responsible for sex determination or they might be situated downstream in the sex determination cascade. Many of these genes do not fulfill a sex-related function in crustacean *M. nipponense* [124] and in crustaceans *D. magna* and *F. chinensis*, in the latter of which only *dsx* and *Tra* play a role in sex determination (see 3.2.2.2.). Moreover, upstream sex-determining genes are known to be poorly conserved. Another unlikely, but possible situation could be that these genes are sex-determining, but were not selected by BSA because the scaffold on which the gene was situated carried less than five SNPs (section 3.3.3.) That is why the Arthropod sex-determining gene homologs in *Artemia* should be further investigated.

3.5.5. *Artemia* candidate sex-determining genes

Similarly to most arthropods, sex in *Artemia* is determined genetically. Even though *Daphnia* is the closest related organism to *Artemia* of which the environmental sex-determining system and gene switch *dsx* are known, *Artemi*a does not have an environmental sex determination system. It is known that regulatory networks that control sex determination vary tremendously between and even within animal species and are among the fastest evolving biological networks [178].

Like most high-value crustaceans, such as *L. vannamei*, *P. monodon* and *Macrobrachium*, and like the silk moth *Bombyx mori*, *Artemia franciscana* has a WZ/ZZ sex-determining system. Sex-determining genes in arthropod WZ/ZZ sex-determining systems have not been identified yet and currently, only candidate sex-determining genes are known in *B. mori* [62] and in *Macrobrachium nipponense* [124].

On a short term, further research could be focused on three genes selected by BSA that showed mostly heterozygous SNPs in *Artemia* females and homozygous SNPs in males and were closely connected with primary sex determination in arthropods. *Cytochrome P450* represented the most valid candidate sex-determining gene, because it had only non-synonymous SNPs, a sex-specific secondary protein structure and has already been put forward as a candidate sex-determining gene in crustacean *M. nipponense* through transcriptomic evidence. The gene containing CRAL-TRIO domain was highly homologous with *SEC14* that influences progeny gender in *Daphnia* and contained non-synonymous SNPs, causing a sex-specific secondary protein structure in male and female *Artemia*. *Zinc finger, C2H2*, although selected by BSA, had no non-synonymous SNPs in the exonic region, but is one of two candidates for primary sex determination in *B. mori* (WZ/ZZ), so it could be investigated as well.

In order to narrow down the gene selection to one candidate sex-determining gene, more research is needed involving SNPs in regulatory regions, which were not detected with our methods testing only SNPs in the exonic region of each gene. Indels detected, but not analyzed in this work could be included in a future BSA analysis. Candidate sex-determining genes without functional description could be analyzed after the planned improvement of the genome annotation and manual curation of all the genes thought to be involved in sex determination. Finally, arthropod sex-determining genes found in the *Artemia* genome, which were not selected as candidate sex-determining genes, should be analyzed in-depth. Perspectives for future research are discussed in more detail in Chapter 5.

Chapter 4

The *Artemia* genome

4.1. Abstract

Next-generation sequences of *Artemia franciscana* were assembled *de novo.* The 1,310 Mbp genome sequence (N50 = 14,784 bp; GC content = 35% and 176,667 scaffolds) was annotated, predicting 188,101 genes with an average length of 692 bp. Ninety percent of the *Artemia* expressed sequence tags (ESTs) available at the National Center for Biotechnology Information (NCBI), as well as 92.2% of the RNAseq reads from cysts at different metabolic stages (anoxia; diapause; quiescence and hydration) and from nauplii kept at different salinities (30 g/l and 200 g/l) were present in the *Artemia* genome, revealing that the functional part of the genome under the studied sampling conditions wass virtually fully represented in the assembly. Genome finishing strategies are however still needed to reduce fragmentation of the genome.

4.2. Introduction

The crustaceans represent a diverse group of 66,914 species [236] of which currently only two publicly available genome sequences are completed at the level of contigs or scaffolds: the annotated genome of the branchiopod water flea *Daphnia pulex* [39], and a draft assembly of the salmon louse genome *Lepeophtheirus salmonis* [149] of which the annotation is still in progress [60,85].

Artemia species*,* also known as brine shrimps (Pancrustacea, Branchiopoda, Anostraca), are cosmopolitan planktonic crustaceans inhabiting hyper-saline ecosystems and salt works on all continents, except Antarctica. Their geographical dispersion explains the high diversity within *Artemia*. In addition to a hyper-saline environment that is subject to considerable salinity shifts, various other factors, such as high UV radiation, large differences in temperature, desiccation and anoxia represent stresses in the *Artemia* habitat [66]. Being extremophiles, *Artemia* species have adapted to these harsh and unstable conditions with a great plasticity to abiotic environmental fluctuations and by the production of encysted embryos (cysts) under stressful conditions [89,104]. Cysts can be stored for years. In research, this allows “common garden” experiments with cysts of different generations [65]. The existence and availability of cysts have assured the importance of brine shrimps in many biological disciplines. *Artemia,* also called “the aquatic *Drosophila*” [64], is used for toxicity testing of compounds, water detoxification, and as a model for crustacean metabolism, evolution and disease research [173,190]. The diploid *A. franciscana* (2n = 42) has a relatively small 0.93-Gb genome (Chapter 2), considering that the average crustacean genome size is over four times larger (section 1.2.3.) [173]. Here, the first *Artemia de novo* genome sequence is presented and analyzed.

4.3. Materials & Methods

4.3.1. Samples for sequencing

4.3.1.1. PE and MP DNA

Samples for paired-end (PE) and Mate-Pair (MP) DNA sequencing were taken as follows. Cyst material of *A. franciscana* strains from San Francisco Bay (SFB) and Vinh Chau (VC) was hatched and reared until sexual maturity as described in section 2.3.1.

Three controlled crosses (C1, C2 and C3) between VC (♀) and SFB (♂) were made, each resulting in F_1 progeny that was collected, grown until maturity, rinsed and sexed. For gut evacuation before DNA extraction, the offspring were kept overnight in a cellulose solution (1.5 g/l; Sigma, type 20) [41], followed by removal of the brood pouch in females. Progeny were stored individually at -20 °C.

4.3.1.2. RNA

-1) Parthenogenetic Artemia

Cysts in diapause

Cysts from parthenogenetic *Artemia* (strain Tuz Lake, Kazakhstan, ARC1761), known to have a high percentage of cysts in diapause, were kept in a brine solution. Hatching percentage was tested with and without freezing to break diapause. The Tuz Lake strain was sampled once.

-2) Artemia franciscana

An overview of *A. franciscana* samples taken for RNAseq is provided in Figure 4-1. More details about sampling methods are described below.

Quiescent cysts

Before the three sampling rounds, cyst metabolic stages were synchronized by hydration in water and subsequent dehydration in brine. The quiescent cysts were sampled once. Each experiment in the following protocols was done in triplicate.

Hydrated cysts

Cysts were taken out of the brine, rinsed and hydrated with tap water for 1 h with aeration, sieved and rinsed with mQ. Samples were taken with a small spatula cleaned with ethanol and put in labeled eppendorfs, previously put in liquid N_2 to flash-freeze.

The remainder of the hydrated cysts was put in (1) regular, non-autoclaved sea water (30 g/l) in a cone in the dark at 28°C and left to hatch for 24 h with aeration or (2) left to hydrate longer for a total of 6 h with aeration, then used for testing under anoxic conditions.

Instar I larvi, hatched in the dark, kept at low salinity

Nauplii were separated from cysts by removing aeration for a few minutes and siphoning nauplii to a new, clean, aerated cone. Samples were taken with a small spatula after 15 and 45 min in regular sea water by putting half a cone of nauplii over a sieve per sample time and rinsing with mQ. Samples were flash-frozen with liquid N_2. Between sampling times, the sieve was washed with tap water and rinsed with autoclaved mQ

Instar I larvi, hatched in light, kept at low salinity

Nauplii were separated from cysts by removing aeration for a few minutes, pipetting the nauplii (25 ml pipette) over a sieve and adding them to a new, clean, aerated cone with Instant Ocean 30g/l. Samples were taken after 15, 30, 45 and 60 min in new 30 g/l Instant Ocean by putting 50 ml (with 25 ml pipette) from a cone of nauplii over a sieve per sample time, rinsing with mQ over the sieve and taking 100 µl of mQ/*Artemia* mix per sample with a 1-ml pipette with cut tips. Samples were flash-frozen with liquid N_2. Between sampling times, the sieve was washed with tap water and rinsed with autoclaved mQ.

To check the cyst/nauplius ratio in the samples (inevitably, some cysts or cyst shells cling to the nauplii) and the number of animals per sample, we observed 1 ml of the 50-ml samples. It contained 93 Instar I nauplii/ml and 9 cysts (full or empty)/ml, or in other words, each sample contained approximately 10% cysts (full or empty).

Instar I larvi, hatched in light, kept at high salinity

Nauplii were separated from cysts by removing aeration for a few minutes, pipetting the nauplii (25-ml pipette) over a sieve and adding them to a new, clean, aerated cone with Instant Ocean 200g/l. Samples were taken after 15, 30, 45 and 60 min in new 200 g/l Instant Ocean by putting 50 ml (25 ml-pipette) from a cone of nauplii over a sieve per sample time, rinsing with mQ the sieve, and taking 100 µl of mQ/*Artemia* mix per sample with a 1 ml-pipette with cut tips. Samples were flash-frozen with liquid N_2. Between sampling times, the sieve was washed with tap water and rinsed with autoclaved mQ.

Anoxic cysts

Two scintillation vials with hydrated cysts were put under anoxic conditions in the light on a rotor. Anoxic conditions were reached according to a modified protocol [36]. In scintillation glass vials of 8 ml, 70-80 mg dry cysts were placed. N_2 gas was flown through the vial to remove the air in the vial, and then the vial was capped and set aside, allowing the N_2 to seep into the porous parts of the outer cyst shell. N_2 was bubbled through buffer (0.25M NaCl in 0.05M phosphate buffer, pH 7.2) during a minimum of 4 h. With the N_2 still bubbling, 6 ml of the anoxic solution was transferred into each vial without any air leaking into the vial and Parafilm was wrapped around the tightly shut vial cap. Vials were oscillated on a rotating platform shaker (50 revolutions/min) for the first 24 h to hydrate the cysts completely and to ascertain that even traces of oxygen (if present) are consumed. The vials containing the anoxic, hydrated cysts were stored at room temperature and in ambient light. Samples were taken after 25 h of anoxia (1 hour after metabolic activity stops). The content of each vial was sieved and rinsed with mQ. Per sample, 100 µl of mQ/*Artemia* mix was taken with a 1-ml pipette with cut tips. Samples were flash-frozen with liquid N_2. Between sampling times, the sieve was washed with tap water and rinsed with autoclaved mQ.

The anoxic state in the scintillation vials was verified indirectly by observing the cysts under a binocular microscope after three days of anoxia. In a sample of 500 cysts, there was one clear umbrella and a few (2-3) cysts with possible breaking. In similar conditions, but with oxygen, the same sample had a hatching percentage of 70% after 24 h. Crushing the cysts between microscope slides to push the non-hatched embryos out of their cyst shells was done to be sure that practically all cysts contained an embryo and that the results were not biased by an excessive amount of empty cyst shells.

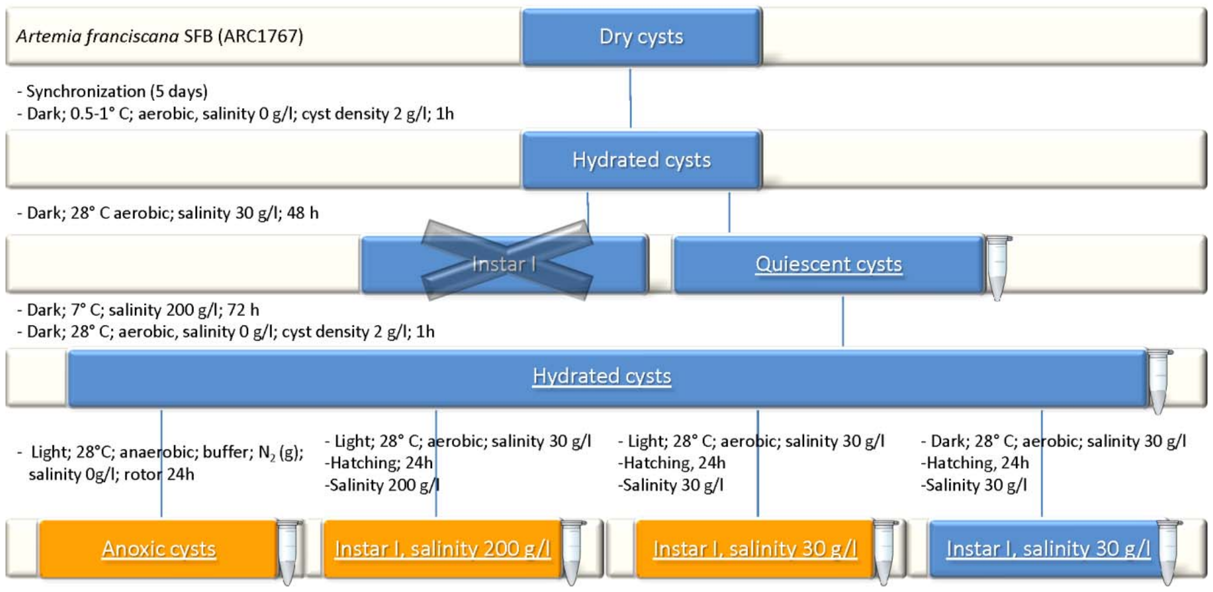

Artemia franciscana SFB (ARC1767)
Dry cysts
- Synchronization (5 days)
- Dark; 0.5-1° C; aerobic, salinity 0 g/l; cyst density 2 g/l; 1h
Hydrated cysts
- Dark; 28° C aerobic; salinity 30 g/l; 48 h
Instar I
Quiescent cysts
- Dark; 7° C; salinity 200 g/l; 72 h
- Dark; 28° C; aerobic, salinity 0 g/l; cyst density 2 g/l; 1h
Hydrated cysts
- Light; 28°C; anaerobic; buffer; N2 (g); salinity 0g/l; rotor 24h
- Light; 28° C; aerobic; salinity 30 g/l
-Hatching; 24h
-Salinity 200 g/l
- Light; 28° C; aerobic; salinity 30 g/l
-Hatching, 24h
-Salinity 30 g/l
- Dark; 28° C; aerobic; salinity 30 g/l
-Hatching, 24h
-Salinity 30 g/l
Anoxic cysts
Instar I, salinity 200 g/l
Instar I, salinity 30 g/l
Instar I, salinity 30 g/l

Figure 4-1. Sampling (each sampling step is underlined) of quiescent, hydrated and anoxic cysts, Instar I larvae hatched in light and kept at salinities 30 g/l or 200 g/l and Instar I larvae hatched in the dark and kept at a salinity of 30 g/l. All samples were taken from *A. franciscana*, strain SFB (USA).

Artemia sampled under seven different conditions including different life cycle stages were analyzed (Table 4-1).

Table 4-1. *Artemia* sampling for RNAseq: treatment, life cycle stage, sampling time after the start of the treatment, number of biological replicates for each sampling time, cyst ID from the *Artemia* Reference Center cyst bank, species and strains used are presented.

Treatment	Life cycle stage	Sampling time (min)	Biological replicates	Cyst ID	Species	Strain
diapaused	cysts	0	1	ARC1761	Parthenogenetic *Artemia*	Tuz Lake
quiescent	cysts	0	1	ARC1767	*A. franciscana*	SFB
hydrated	cysts	0	3	ARC1767	*A. franciscana*	SFB
anoxic	cysts	60	3	ARC1767	*A. franciscana*	SFB
hatched in light, salinity 200 g/l	Instar I larvi	15; 30; 45; 60	3	ARC1767	*A. franciscana*	SFB
hatched in light, salinity 30 g/l	Instar I larvi	15; 30; 45; 60	3	ARC1767	*A. franciscana*	SFB
hatched in the dark, salinity 30 g/l	Instar I larvi	15; 45	3	ARC1767	*A. franciscana*	SFB

4.3.2. Sequencing

4.3.2.1. PE DNA

DNA was extracted as described in section 2.3.2. from 120 F_1 progeny obtained from cross C1. DNA quality and yield of each individual DNA sample were assessed with a NanoDrop ND 1000 spectrophotometer (Thermo Scientific) [48] and samples were diluted to a concentration of 30 ng/µl. A “male” and a “female” pool of 55 male and 65 female F_1 DNA samples respectively, were made by bulking equimolar amounts of DNA. Illumina DNA sequencing libraries of insert sizes 200 and 500 bp were prepared from each pool (1-5 µg DNA) with the "Illumina TruSeq DNA Sample Preparation Kit" standard protocol and sequenced by Fasteris (http://www.fasteris.com). DNA fragmentation (50-500 bp), end-repair, A-tailing and adapter ligation were followed by gel isolation of fragments of the required insert size and PCR amplification. The libraries were quantified, diluted to 10 nM and analyzed in a 2x100 bp run on an Illumina HiSeq 2000 instrument. The PE reads were used to assemble and scaffold the *Artemia* genome.

4.3.2.2. Cre-lox MP DNA

DNA was extracted as described in section 2.3.2 from one pool of 50 C2 F_1 progeny and one pool of 50 C3 F_1 progeny, each pool containing 25 males and 25 females. DNA quality and yield were assessed with a Quant-iT™ dsDNA High-Sensitivity Assay Kit (Invitrogen). To meet the sample requirements of the sequencing facility, 3 µg of DNA of each pool was combined into one pool (final DNA concentration 30 ng/µl).

From the bulked *Artemia* DNA, libraries with an insert size of 3 kb were prepared (NXTGNT Laboratory of Pharmaceutical Biotechnology, Belgium). Cre-lox MP sequencing [210] was applied to the libraries on an Illumina HiSeq 2000 instrument (Scripps Research Institute, USA). The MP reads were used to scaffold the *Artemia* genome.

4.3.2.3.PE RNA

For each of the 38 samples, RNA was extracted by a combined method with TRIzol (Invitrogen), the RNeasy Mini Kit (Qiagen) and the DNase I Kit (Qiagen) [206]. Briefly, each freeze-dried, homogenized sample was supplemented with 500 µl of TRIzol. After centrifugation, the supernatant was incubated (5 min at 23°C) and 100 µl of chloroform was added. The mix was incubated again (2 min at 23°C) and centrifuged. The clear supernatant was mixed with 300 µl of 70% ethanol. The following steps were done with the RNeasy Mini Kit.

The mix was transferred to an Rneasy Mini spin column, it was centrifuged and the flow-through was discarded. To the column, 350 µl of buffer RW1[12] was added, and the flow-through was discarded after centrifugation. A mix of 10 µl DNase I and 70 µl of buffer RDD[13] were added to the column and incubated (30 min at 23°C). Then 350 µl of buffer RW1 was added to the column and after incubation (5 min at 23°C) and centrifugation, the flow-through was discarded. After addition of 500 µl of buffer RPE to the column and after centrifugation, the flow-through was discarded once more. This last step was repeated. Finally, the column was dried, 30 µl of RNase-free water was added, and it was incubated (1 min at 23°C) and centrifuged to elute the RNA. RNA quality and yield of each sample were assessed with the Agilent 2100 bioanalyzer (Agilent Technologies, USA). Samples were diluted to an average of 40 ng RNA/µl.

The 38 RNA sample libraries were prepared by subsequent use of the TruSeq™ mRNA enrichment and TruSeq™ RNA sample preparation kits (Illumina) and paired-end sequenced at the Genomics Core facility (UZ Leuven, http://gc.uzleuven.be) on an Illumina HiSeq 2000 instrument (2x100 bp, unstranded, insert size 200). The sequencing data were quality-trimmed with CLC Assembly Cell 4.06 software (CLC bio, http://www.clcbio.com/products/clc-assembly-cell). The RNAseq reads were used for correction and annotation of the *Artemia* draft genome sequence.

4.3.3. Genome assembly

An overview of the most important assembly steps is provided in Figure 4-2. Overrepresented PE reads (> 0.6% of the library) that contained primer or adapter sequences were removed. PE and deloxed MP reads were quality-trimmed (quality score minimum 20, minimum length of PE and MP reads, 65 bp and 36 bp, respectively). PE reads were assembled into a draft genome, with the trimmed PE and MP reads set as scaffolding information by means of the software CLC Assembly Cell 4.06 [35]. As *Artemia* is known for its many repetitive regions [26], creating a delicate balance between assembling and collapsing too many similar regions, we opted for the conservative (non-greedy) assembly approach of CLC. An adjusted insert size per library (interval within which 99% of the paired reads of the library map) and adjusted word and bubble sizes (64 and 100, respectively) were used to improve the assembly quality.

With the software SSPACE 2.0 [19], the draft genome was scaffolded[14] with all trimmed paired PE and MP reads.

---

[12] Rneasy Mini Kit

[13] Rneasy Mini Kit

[14] SSPACE 2.0 parameters:
-x 0 -g 3 -k 3

Short scaffolds (≤ 1kb) were mapped with default parameters and a mismatch cost of 3 onto long scaffolds (≥ 1kb) with CLC Assembly Cell 4.06 (parameters -x 3). Short scaffolds mapping with a minimum similarity of 0.8 were considered redundant and removed from the assembly.

Next, scaffolds of mitochondrial origin were identified by BLAST (blastn, E-value 1E-10) of the *A. franciscana* mtDNA Sanger sequence (Genbank X69067.1 [208]) onto the assembly and were removed, retaining only nuclear genome sequences.

Trimmed PE and MP reads with one sequence mapping on a gap (NNN) and its pair mapping on an ungapped scaffold region as well as those connecting two scaffolds were selected as a subset read database for re-scaffolding the draft genome with SSPACE 2.0 using the same parameters.

Subsequently, gaps created by scaffolding were filled[15] in by four iterations with GapFiller 1.9. [20].

The trimmed *Artemia* RNAseq data was mapped onto the assembly by means of CLC Assembly Cell 4.0.13 (parameters -x 3), with an insert size of 4 kb to bridge the distance between exons in the genome. A consensus sequence between the assembly and the trimmed RNAseq data was obtained with CLC Assembly Cell 4.0.13 (clc_find_variations) to improve the subsequent annotation.

Scaffolds containing one or more sex markers (Chapter 3) were anchored in the order indicated by the *Artemia* high-density genetic linkage map (Chapter 2), with orientation data only applicable to scaffolds containing more than one sex marker. Anchoring was performed with CLC Assembly cell 4.0.13 using a golden path (AGP) file provided in Table 3-1.

---

[15] GapFiller 1.9. parameters:
-m 30 -o 2 -r 0.7 -n 10 -d 50 -t 10 -g 0 -i 4

Finally, scaffolds originating from contaminant microbial, fungal or viral organisms were eliminated from the assembly after performing BLAST (blastx) on the Refseq protein databases “microbial”, “fungi”, “viral” and “invertebrate” (release 56 [150]), removing only scaffolds containing 100% non-invertebrate protein with a minimum ID of 70% and a maximum length of 5kb without contaminant hits, to avoid eliminating scaffolds containing unknown arthropod genes. Short scaffolds (≤ 1kb) still present in the assembly at this point were removed.

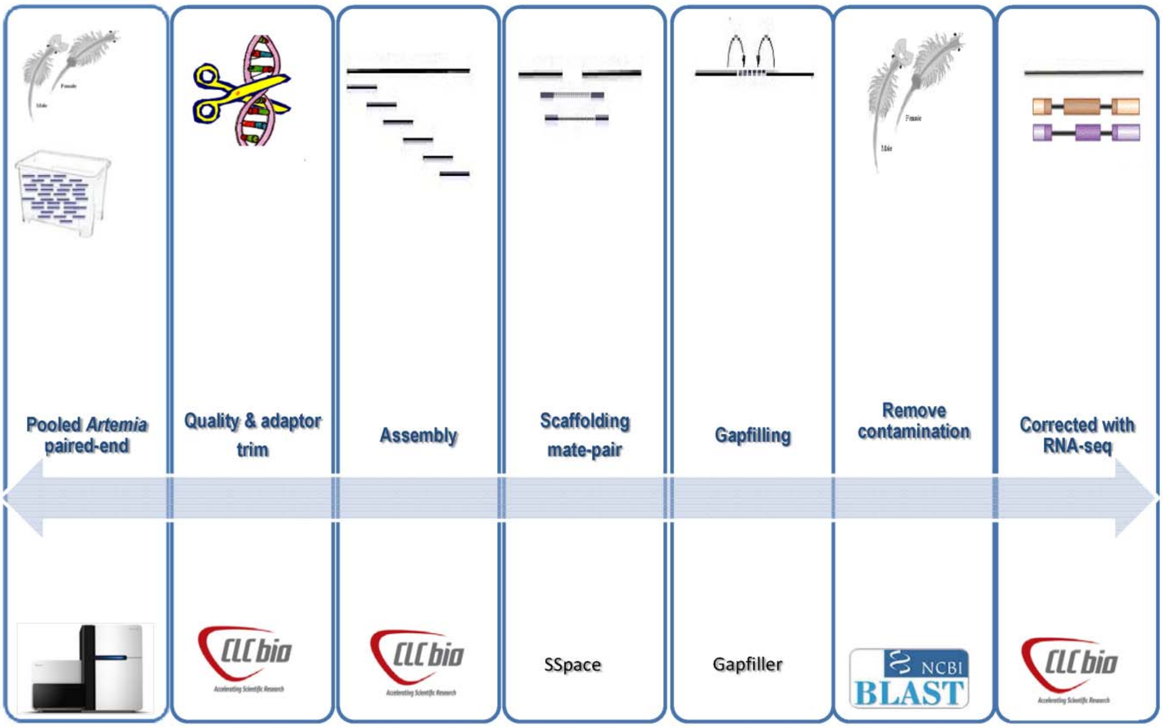


Figure 4-2. Assembly of the *Artemia* draft genome sequence. Assembly steps are shown from left to right: Paired-end sequencing of pooled *Artemia*, quality and adaptor trimming, assembly, scaffolding with mate-pair sequences, gapfilling, removing of contamination, correcting with RNA-seq. The software used for each step is shown below.

The completeness of the assembly was evaluated by applying BLAST (blastn, 1E-3) of the 34,618 known *Artemia* ESTs (NCBI) to the assembly and by mapping of the RNAseq reads onto the *Artemia* genome assembly.

4.3.4. Genome annotation

The 176,667 genome fragments (scaffolds), of a total length of 1,310,312,779 bp were annotated. *De novo* repeat families were identified and modeled with RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html).

To insure only real repeats were present in the consensus repeat database, the repeats were hard-masked with RepeatMasker [198] using a library of known repeats in spider mite [72], and repeat length in the consensus repeat database was determined.

Splice junctions on the scaffolds were discovered with TopHat[16] 2.0.8b [203] to map the trimmed *Artemia* RNAseq data onto the selected *Artemia* scaffolds and subsequently analyzed with the SpliceMachine signal sensor components [17][46], trained specifically on *A. franciscana* data. Gene evidence, such as the splice junctions (min. 10 junctions per intron), RNAseq mapping and the EST and protein BLAST results were integrated (Figure 4-3) into a structural annotation by the coding gene prediction platform EuGène 4.1.[18] [57].

Functional annotation was done with Refseq “invertebrate” protein [162] and invertebrate EST BLAST results via the online annotation platform ORCAE [195]. Gene models were made available online for partners, for manual curation on the ORCAE platform with GenomeView [5].

---

[16] Single-end reads after trimming:

```
tophat -p 4 --read-realign-edit-dist 0 --library-type fr-unstranded
```

Paired-end reads after trimming:

```
tophat -p 4 --read-realign-edit-dist 0 --mate-std-dev 50 --library-type fr-unstranded
```

[17] SpliceMachine parameters:

```
SMachine.cmd                "splicemachine.pl "
SMachine.isScaled           0
SMachine.accP               1.768521
SMachine.accB               3.233744
SMachine.donP               2.065017
SMachine.donB               3.380102
SMachine.startP             0.052
SMachine.startB             0.308
SMachine.tSpliceB           0.0
#SMachine.format            GFF3
```

[18] Eugène parameters:

```
eugene -p gl -E -B
```

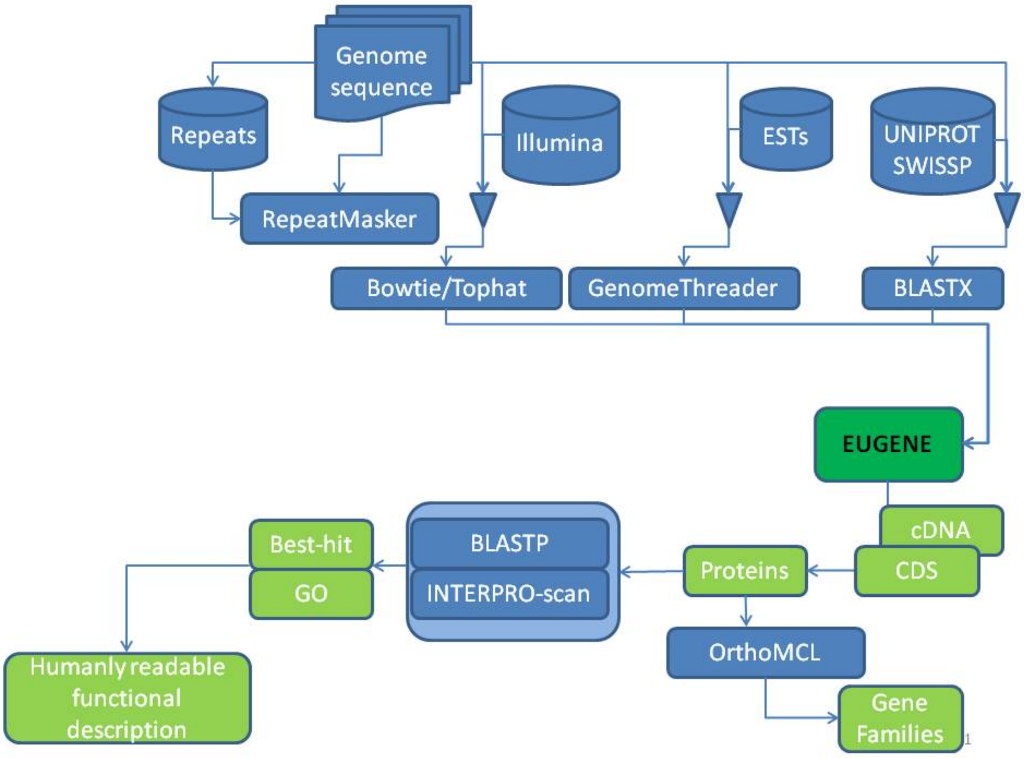


Figure 4-3 Annotation pipeline for the *Artemia* genome. Repeat families were identified and modeled with RepeatModeler and repeats were masked with RepeatMasker, using a repeat library of spider mite [72]. Gene evidence such as the RNAseq mapping, the EST and protein BLAST results and the Illumina PE reads were combined into a structural annotation by the gene prediction platform EuGène. BLAST was performed on all structures identified as proteins with Refseq "invertebrate" protein (blastp) and invertebrate ESTs (blastn), whereas gene families and domains were identified with InterProScan [231].

To evaluate the completeness of the annotated protein-coding genes, BLAST (blastp, E-value 0.1) of the protein sequence from the annotated genes to the NCBI "nr protein" database was performed and bins of the length ratio between the predicted genes and the best BLAST hit were produced. Predicted genes approaching full length were expected to have a length ratio of [0.9,1.1], whereas too short or too long genes were expected to have a length ratio of [0.0,0.9[and over 1.1, respectively.

4.3.5. Homology of *L. vannamei* protein in the *Artemia* genome

To determine whether *Artemia* could be a useful model for high-value crustaceans such as *L. vannamei*, BLAST (tblastn, E-value 1E-3) of 512 *L. vannamei* nuclear proteins (NCBI) was performed onto the *Artemia* genome.

4.4. Results

From the male and female bulk, 814 and 969 million (M) paired-end DNA reads of 100 bp were generated, respectively, resulting in an initial average coverage of 88X and 104X, (total average coverage of 192X), based on the 0.93-Gb *Artemia* genome size estimated by flow cytometry (Chapter 2).

From the mixed bulk, 264 M deloxed mate-pair reads of 100 bp were generated, reaching an initial average genome coverage of 25X. These reads were used for further scaffolding into longer DNA stretches and for improved bridging of highly repetitive genomic regions.

Paired-end RNAseq resulted in 1,240 M single reads of 100 bp in total. Additional information about the raw data in the paired-end and mate-pair DNA and the RNA libraries is provided in Table 4-2.

Table 4-2. Libraries used for sequencing of DNA (paired-end (PE) and mate-pair (MP)) and RNA (PE). Read length is 100 bp for each library. Library type (DNA/RNA), library name, sequencing type (PE/MP), sampling material (cysts/larvi/males/females), library insert size and the number of untrimmed reads are provided.

Library type	Library name	Sequencing	Sampling material	insert size (bp)	Number of reads
DNA	GLX-17-A7	PE	females	200	196,166,364
	GLX-17-B2	PE	females	200	200,686,680
	GLX-19-A7	PE	females	500	152,693,006
	GLX-19-B7	PE	females	500	214,697,742
	GLX-19-B4	PE	females	500	205,137,108
	GLX-16-A6	PE	males	200	171,548,078
	GLX-16-B1	PE	males	200	176,238,374
	GLX-18-A8	PE	males	500	136,879,890
	GLX-18-B3	PE	males	500	138,005,250
	GLX-25-B2	PE	males	500	191,268,838
	Run40-8-deloxed	MP	males & females	3000	264,135,600
RNA	HSR000185	PE	cysts	200	32,547,084
	HSR000186	PE	cysts	200	25,343,244
	HSR000187	PE	cysts	200	31,329,090
	HSR000188	PE	larvi	200	35,980,948
	HSR000189	PE	larvi	200	22,626,494
	HSR000190	PE	larvi	200	20,303,832
	HSR000191	PE	larvi	200	31,303,490
	HSR000192	PE	larvi	200	22,931,256
	HSR000193	PE	larvi	200	32,946,492
	HSR000194	PE	larvi	200	22,812,268
	HSR000195	PE	larvi	200	31,866,074
	HSR000196	PE	larvi	200	31,232,208
	HSR000197	PE	larvi	200	34,257,294
	HSR000198	PE	cysts	200	24,145,044
	HSR000199	PE	cysts	200	39,949,728
	HSR000200	PE	larvi	200	39,488,216
	HSR000201	PE	larvi	200	30,428,418
	HSR000202	PE	larvi	200	28,414,480
	HSR000203	PE	larvi	200	30,701,874
	HSR000204	PE	larvi	200	40,083,796
	HSR000205	PE	larvi	200	27,162,962
	HSR000206	PE	larvi	200	36,600,120
	HSR000207	PE	larvi	200	40,978,358
	HSR000208	PE	larvi	200	37,938,164
	HSR000209	PE	larvi	200	20,404,532
	HSR000210	PE	cysts	200	23,534,902
	HSR000211	PE	cysts	200	49,794,424
	HSR000212	PE	larvi	200	20,004,244
	HSR000213	PE	larvi	200	22,304,260
	HSR000214	PE	larvi	200	54,073,866

Library type	Library name	Sequencing	Sampling material	insert size (bp)	Number of reads
	HSR000215	PE	larvi	200	27,993,664
	HSR000216	PE	larvi	200	25,947,946
	HSR000217	PE	larvi	200	47,418,580
	HSR000218	PE	larvi	200	52,025,692
	HSR000219	PE	larvi	200	24,152,928
	HSR000220	PE	larvi	200	54,260,774
	HSR000221	PE	larvi	200	23,065,494
	HSR000222	PE	cysts	200	31,355,520

Quality trimming of the raw PE and MP reads removed 13.98% (1,534 M trimmed PE reads left) and 9.65% of the raw reads, respectively, whereas quality trimming of the RNAseq reads eliminated 7.38% of the raw reads.

4.4.1. Genome assembly

To evaluate the different assembly steps, the evolution of the different key assembly values is listed in Figures 4-4; 4-5 and 4-6. More assembly data are available in Table 4-3.

Table 4-3. Overview of key assembly values (number of scaffolds, longest scaffold, average scaffold length, N50, mapped reads (PE), paired reads, mapped nucleotides, corrected average coverage (PE), assembly size and gap %) for the different *Artemia* genome assembly stages.

	First draft genome	Scaffolded with SSPACE	Removed redundancy	Removed mitochondrial genome	Scaffolded again with SSPACE	Gaps filled	Corrected with RNAseq
Scaffolds	991,259	565,240	330,306	330,200	297,498	297,498	297,498
longest scaffold (bp)	65,553	150,497	150,497	150,497	152,656	154,159	154,146
average scaffold length (bp)	1,422.67	2,546.01	4,120.70	4,116.56	4,579.67	4,614.52	4,614.48
N50	4,021	10,681	11,676	11,661	13,697	13,826	13,826
Mapped reads (%)	99.32	99.31	99.48	99.44	99.44	99.50	99.5
Paired reads (%)	52.49	54.16	54.62	54.45	54.49	64.93	64.70
Mapped nucleotides (%)	96.59	96.64	98.27	98.22	98.21	98.71	98.71
Corrected average coverage	122.67	122.98	134.08	134.19	134.20	126.26	126.27
Assembly size (Mb)	1,410	1,439	1,361	1,359	1,362	1,373	1,373
Gaps (%)	18.86	20.64	21.74	21.74	21.93	17.24	17.24

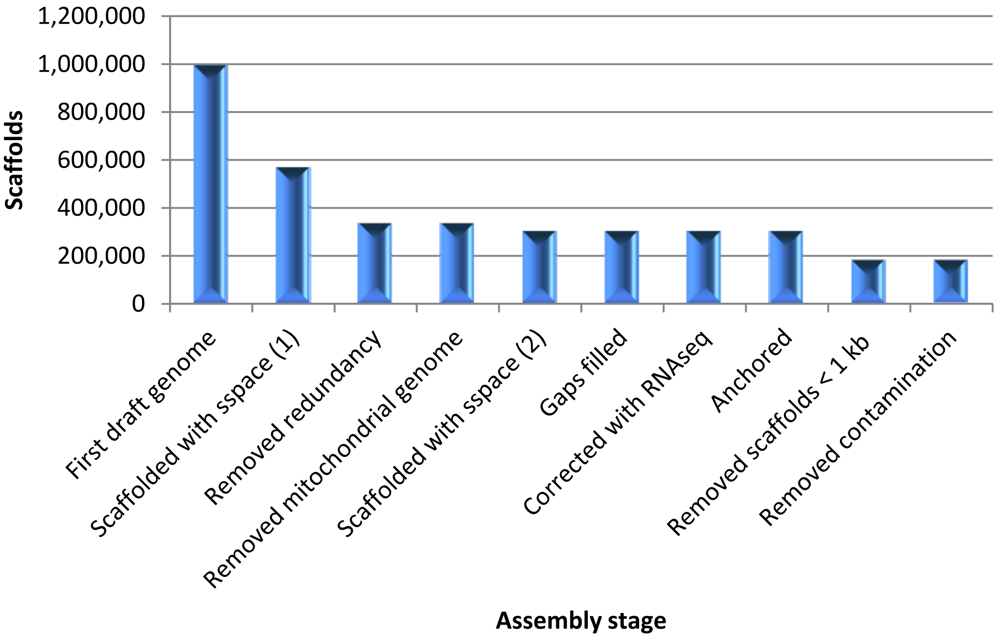


Figure 4-4. Evolution of the number of scaffolds in the subsequent *Artemia* assembly stages.

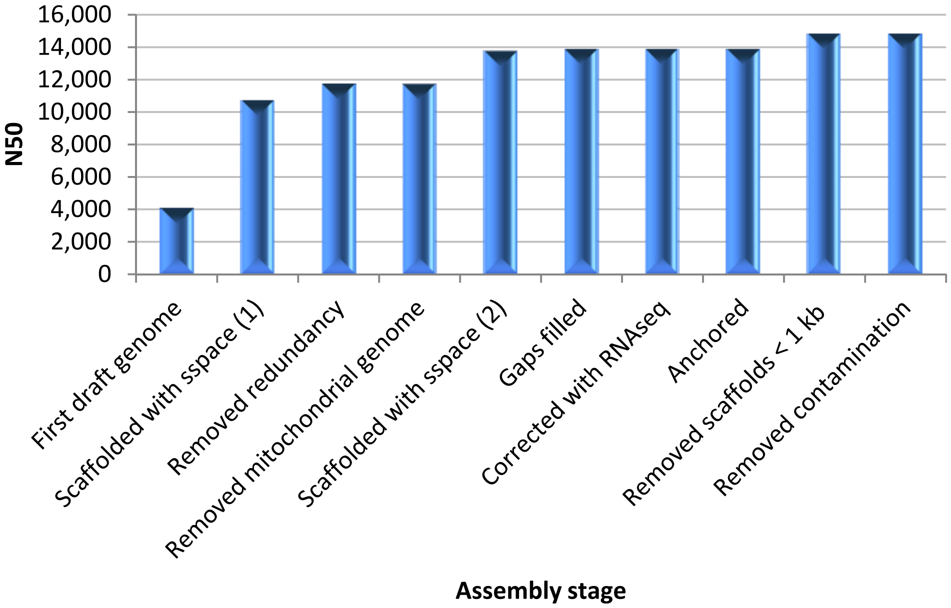


Figure 4-5. Evolution of the N50 value in the subsequent *Artemia* assembly stages.

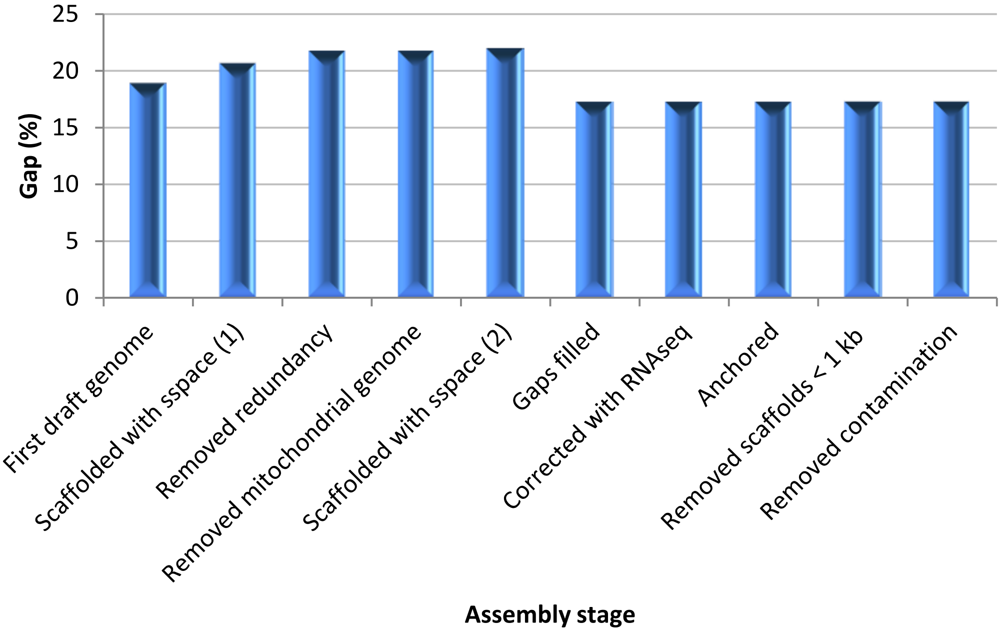


Figure 4-6. Evolution of the gap percentage in the subsequent *Artemia* assembly stages.

The number of scaffolds consistently decreased throughout the assembly steps, whereas the N50 value increased from the first assembly step to the last one. Scaffolding steps particularly elongated and reduced the amount of scaffolds, while increasing the percentage of gap nucleotides. In the genome, 234,934 short scaffolds (≤ 1 kb) were considered redundant, since they mapped to longer scaffolds with a minimum similarity of 0.8.

The mitochondrial genome consisted of 106 scaffolds, of which the longest scaffold showed 97% average similarity with the *A. franciscana* complete mtDNA (Genbank query X69067.1). Finally, gap filling reduced the amount of gap nucleotides (“N”) in the completed nuclear genome assembly from 21.93 to 17.24%. Each gap filling iteration resolved fewer gaps than the previous one and the last iteration resolved less than 1% of the gaps. Mapping of the RNAseq data onto the intermediary genome assembly after gap filling showed 92.2% of the RNAseq nucleotides mapped onto the genome. Of the genome assembly nucleotides, 67.4% were not covered by RNAseq reads, resulting in a corrected average genome coverage by RNAseq of 91X. Correction of the genome assembly with RNAseq data resulted in 402,620 base changes (1,243 from N to A, C, G or T); 20,113 insertions of one or more bases and 38,673 deletions in the consensus sequence. Sixteen scaffolds containing characterized proteins of bacterial origin were removed (Table 4-4).

Table 4-4. Proteins found on contaminant scaffolds removed from the assembly.

Protein	Organism
Putative glycosyl transferase (Glycosyl transferase family 2)	*Rhodospirillum photometricum* DSM 122
cell wall-associated hydrolase	*Vibrio parahaemolyticus* 16
ribosomal protein S15	*Roseovarius nubinhibens* ISM
relaxase/mobilization nuclease domain protein, partial	*Vibrio cholerae* CP1035(8)
bacterial mobilization family protein	*Vibrio cholerae* CP1035(8)
capsid protein F	*Escherichia coli* 97.0003
replication-associated protein A, partial	*Escherichia coli* PA28

As a result, the *Artemia* genome was sequenced at an average coverage of 132X based on the assembled genome sequence (Figure 4-7), producing a final draft assembly of 1,310 Mbp (N50 = 14,784 bp; N80 = 5,071 bp and N90 = 2,772 bp), with a mean GC-content of 35%. The percentage of bases with a specific coverage in the final draft assembly is shown in Figure 4-7.

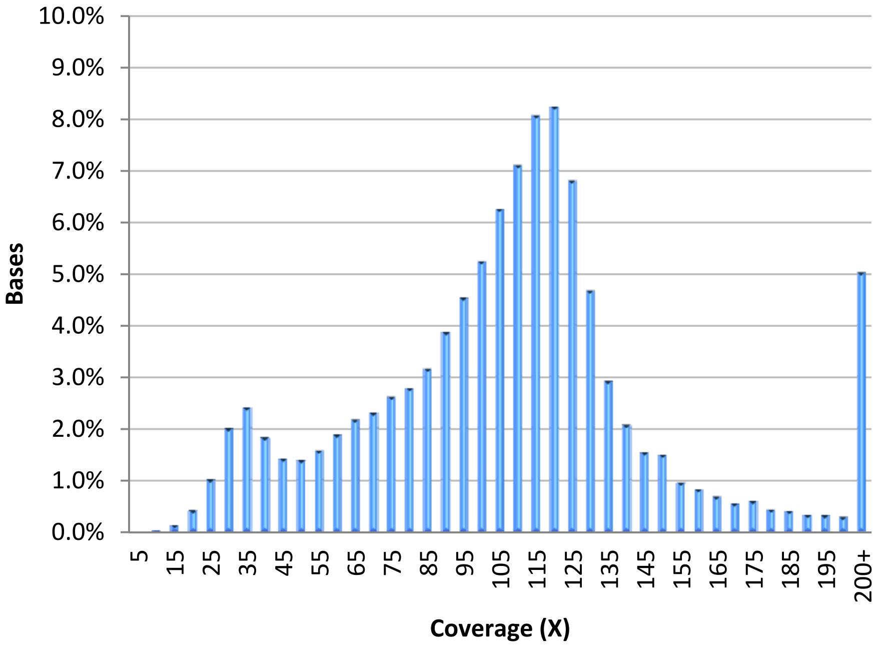


Figure 4-7. Coverage distribution for the assembled *Artemia* draft genome. The trimmed PE reads were aligned onto the draft genome with CLC Assembly Cell (mismatch cost 2). Corrected average coverage per scaffold was used from the alignment.

Another important assembly statistic is the gap percentage (17.3%) or gap length (226 Mb) in the assembly [229].

Of the PE reads, 99% still mapped on the final assembly[19] and more than 64% of the reads were mapped as paired.

Ninety percent of the 34,618 *Artemia* ESTs present on NCBI had BLAST hits in the *Artemia* genome assembly. The shortest scaffold with an EST hit was 1 kb long. However, when considering only scaffolds longer than 10 kb (63% of the genome), only 24% of the same ESTs were detected. These results imply the presence of functional sequences in shorter scaffolds of 1-10 kb. Of all the RNAseq data generated, 92.2% mapped onto the intermediary *Artemia* genome assembly, before it was corrected with the RNAseq data. All 16 tested microsatellites were present in the genome as well (Table 4-5).

Table 4-5. Results from BLAST of 16 *Artemia* microsatellites to the *Artemia* genome assembly.

Scaffold name	*Artemia* species	Bit score	E-value	Microsatellite accession
scaffold_1169	parthenogenetic *Artemia*	140	2.00E-31	EU888846.1
scaffold_13580	parthenogenetic *Artemia*	277	2.00E-72	EU888845.1
scaffold_26246	parthenogenetic *Artemia*	122	8.00E-26	EU888844.1
scaffold_27836	parthenogenetic *Artemia*	141	8.00E-32	EU888843.1
scaffold_19175	parthenogenetic *Artemia*	324	2.00E-86	EU888842.1
scaffold_812	parthenogenetic *Artemia*	554	5.00E-156	EU888841.1
scaffold_154185	*A. franciscana*	791	0	EU888840.1
scaffold_4892	*A. franciscana*	277	3.00E-72	EU888839.1
scaffold_37542	*A. franciscana*	930	0	EU888838.1
scaffold_20470	*A. franciscana*	1101	0	EU888837.1
scaffold_58865	*A. franciscana*	724	0	EU888836.1
scaffold_17234	*A. franciscana*	435	2.00E-120	EU888835.1
scaffold_17245	*A. franciscana*	967	0	EU888834.1
scaffold_53597	*A. franciscana*	764	0	EU888833.1
scaffold_32930	*A. franciscana*	728	0	EU888832.1

4.4.2. Genome annotation

EuGène predicted 188,101 coding genes with an average length of 692 bp, an average exon length of 180 bp, an average of four exons per gene (Figure 4-8) and an average intron length of 253 bp.

---

[19] minimal matching length of the read 50%; minimal similarity of the matching read fraction 80%

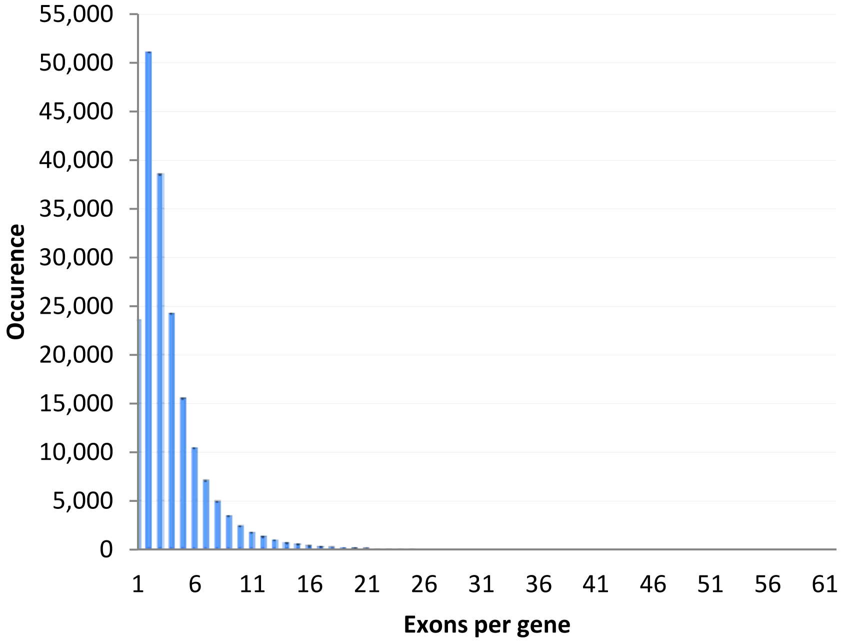


Figure 4-8. Occurrence of exons per gene in the *Artemia* genome

Of the predicted genes, only 1,944 (1%) were full-length (had a length ratio of predicted gene/best BLAST hit between 0.9 and 1.1, Figure 4-9). Most predicted genes did not have a BLAST hit (142,024 genes, comprised in length ratio bin [0.0,0.1[). Genes without BLAST hits indicate either a fragmented or a currently unknown gene. Additionally, the high amount of too short genes shows that there is still fragmentation in the annotated genes.

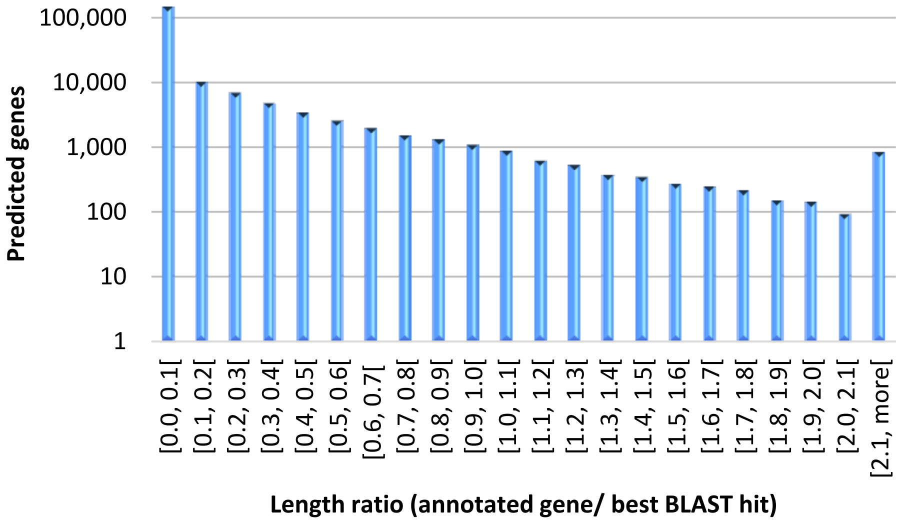


Figure 4-9 Completeness of the predicted coding genes. The number of predicted genes (log scale) for each bin representing a range in length ratio between (1) the length of the annotated gene and (2) the length of the best BLAST (blastp) hit of the predicted gene with the "nr database" from NCBI. Genes with a length ratio between 0.9 and 1.1 are virtually complete.

Of the genome length, 9.9% [20] was exonic, whereas 44.2% consisted of repeats (Figure 4-10).

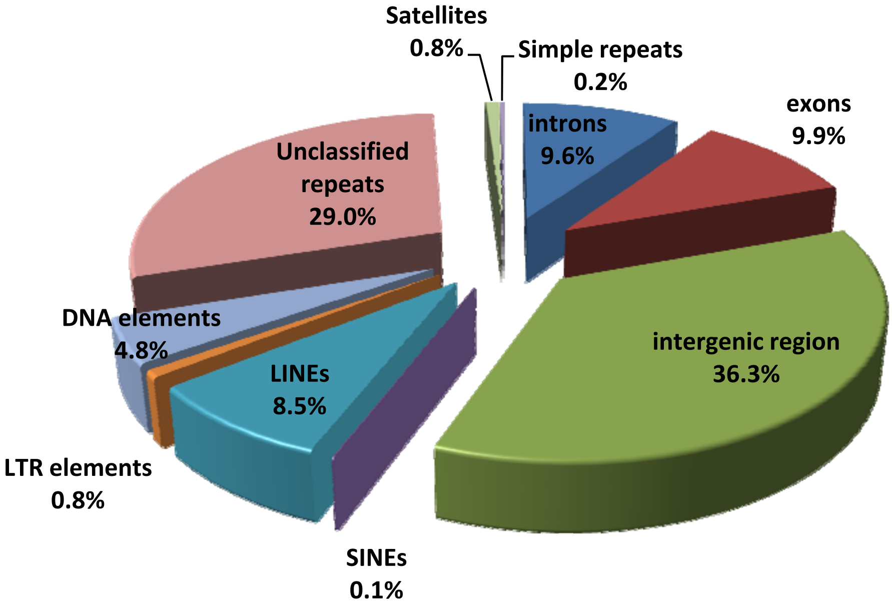


Figure 4-10. Components of the *Artemia* genome, based on component length (bp) compared to the total genome length (bp)

Most of the repeats were interspersed (SINEs; LINEs; LTR elements; DNA elements and unclassified interspersed repeats). Detected LINEs were LINE2 (1.45%) and L3/CR1 (0.69%). The LTR elements only consisted of ERVL (0.01%) and the DNA elements only of hAT Charlie (0.13%). Less than 1% of the genome consisted of satellite repeats[21] (0.79%) and simple repeats (0.21%). Occurrence of repeat lengths in the repeat library is shown in Figure 4-11.

---

[20] The genome was annotated as 9.9% exonic, whereas 33% showed alignment with RNAseq. The reasons for this discrepancy are:

-a small percentage of the RNAseq represent transposable elements

-short scaffolds contain broken genes, which do not end up in the gene prediction due to their short length

[21] Also named short tandem repeats or simple sequence repeats

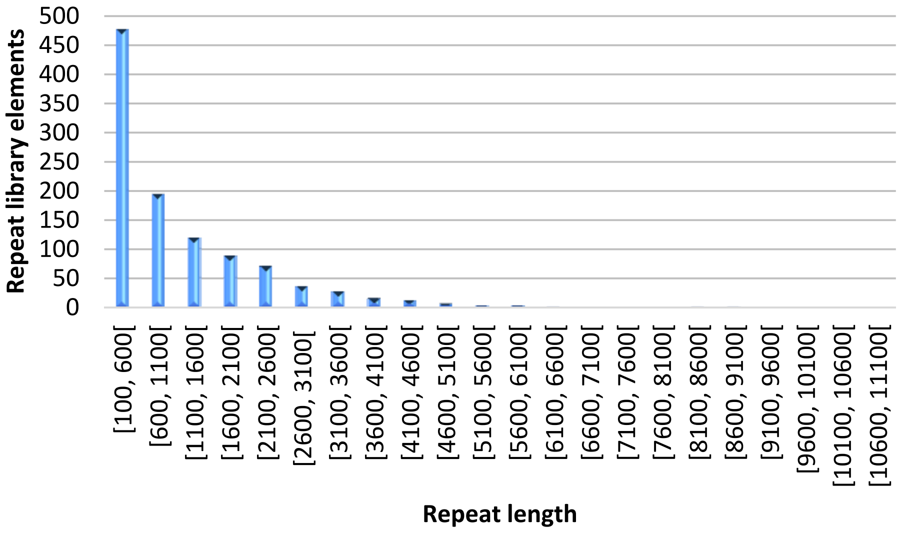


Figure 4-11, Number of repeat library elements with a certain length range.

4.4.3. Homology of *L. vannamei* protein in the *Artemia* genome

Of the 512 *L. vannamei* nuclear proteins available in the NCBI database, 73% were present as homologous proteins (E-value 1E-3) in the *Artemia* genome.

4.5. Discussion

Before the construction of the *Artemia* genome is discussed, it is important to clarify that originally, this assembly was created with PE reads (section 4.3.6.) originating from a bulked segregant analysis (Chapter 3). This initial research purpose is the reason why DNA from pools of full-sib animals was sequenced instead of from a single individual, which is more common for a genome project.

4.5.1. Genome assembly

4.5.1.1. Fragmentation

Since 99% of the PE reads mapped onto the *Artemia* genome assembly, most of the read information was included in the assembly and the sequence that was removed during the assembly process was largely redundant. The ratio of assembly size and estimated *A. franciscana* genome size is however 1,310/930 = 141%, whereas 90-100% is expected [229], implying that 380 Mb of the assembly is still redundant. The low EST hit percentage when only long scaffolds (> 10 kb) are considered means that many genes are still located on the shorter scaffolds (1-10 kb). Many shorter scaffolds are consequently not redundant in the genome.

Based on AFLP fragments from the same cross with different individuals and their 120 F_1 progeny, the average *A. franciscana* nucleotide diversity was estimated at 6.2% (see 2.4.1.). The assembly of sequences from a highly heterozygous diploid genome from one individual remains a known challenge due to differences between homologous chromosomes. Hence, considerably more raw sequence data and auxiliary analyses are required than for assembly of a homozygous or haploid genome [108]. The challenge increases when pooled diploid individuals are used, as in this study. SNPs and indels provoke unresolved bubble formation in the De Bruijn graph [112], the graph type used by the assembly strategy in CLC Assembly Cell. Unresolved bubbles cause sequences that should be assembled to break apart. As shown in Chapter 3, a very large amount of SNPs and indels is found in the PE reads used to assemble the *Artemia genome*, confounding the genome assembly more than could be resolved by subsequent scaffolding with PE and cre-lox MP reads.

Redundancy and fragmentation in the *Artemia* assembly are also partly due to the high content of repetitive elements in the genome, combined with the use of NGS reads, which are unable to span genomic repeats longer than the used read insert sizes (PE: 200 bp and 500 bp; MP: 3000 bp).

4.5.1.3. Repeats

In repeat-rich genomes, the assembly will be broken at each point where a repeat cannot be assembled, leading to a fragmented assembly with short contigs (Figure 4-12) [77].

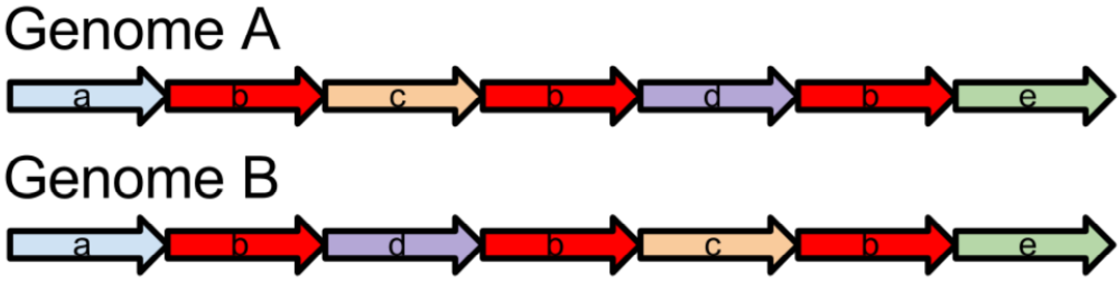


Figure 4-12. Influence of repeats on genome assembly structure. Two possible genomic architectures [47], Genome A and Genome B, each with a repeat b. Both putative genomes can generate the same set of PE reads. With the PE read information, the assembler can identify contigs a, c, d, e and repeat b, but not the entire true genome.

Experimental assemblies of the large and complex human genome from NGS reads show that in *de novo* assemblies, 420.2 Mb of common repeats and 99.1% of validated duplicated sequences are missing from the genome compared to the reference genome assembled with long Sanger reads [9]. NGS may be cheaper and faster than other sequencing techniques, but it has some shortcomings.

The *Artemia* genome consists for 44.2% of assembled repeats. When a genome is assembled with NGS reads, the number of assembled repeats will always be an underestimation of the real number of repeats present in the genome. Any WGS (whole-genome sequencing)-based *de novo* sequence assembly algorithm will collapse identical repeats, resulting in reduced or lost genomic complexity. Recently formed repeats (in humans: microsatellites; Alu and LINE1) have a low level of uniqueness, and, hence, are difficult to assemble [9]. Recent repeats may be assembled because they were bridged by MP reads and elongated by gap filling usually resulting in assembled repeat sequences with only an approximately correct length and structure. Repeats that have been present in a genome for a long time and have become “fixed” in the genome (in humans: retrotransposon-derived repeats) have acquired structural uniqueness over time that make them easier to assemble, and will be the best represented repeats in NGS genome assembly annotations.

4.5.1.3. Completeness of functional sequence

First, assembly statistics will be considered to evaluate completeness and contiguity of the *Artemia* genome assembly. To qualify for annotation, at least 50% of the genes in a genome assembly should be complete, meaning the scaffold N50 value should be close to the median gene length, according to Yandell and Ence [229]. In *Artemia*, the median gene length is estimated to be 5-6 kb (derived from the *Artemia* genome size according to the relationship between genome size and gene length [229]). The assembly shows a N80 of 5,071 bp, close to the median gene length, denoting that approximately 80% of the genes in the assembly should be complete. As the non-greedy assembly algorithm of CLC Assembly Cell was applied, overestimation of N80 should not be an issue.

Another, more hands-on approach to assess the completeness of functional sequences in the assembly is comparison with available *A. franciscana* expression data. Of all the *Artemia* ESTs available on NCBI, 90% mapped on the assembled genome, and 92.2% of all the RNAseq data generated in this study mapped onto the intermediary *Artemia* genome assembly before it was corrected with RNA-seq data. Even though the amount of genome scaffolds is high, these percentages indicate that the functional part of the genome under the studied sampling conditions is virtually complete, meaning that the *Artemia* draft genome can be used for qualitative gene studies.

In spite of N80 and expression data results indicating 80% full-length genes, EuGène predicted 188,101 coding genes of which only 1,944 were full-length. This shows that N80 and expression data alone are not sufficient to evaluate an assembly. To further assess genome completeness, the percentage of most conserved core eukaryotic genes in the genome could be evaluated with CEGMA [156]. However, the core eukaryotic gene database is based on a limited number of genomes and conserved genes are known to be easier to assemble. For these reasons, CEGMA will overestimate genome completeness although it can give a relative estimation of eukaryotic genes in the *Artemia* genome that may be compared with CEGMA results from the genomes of other eukaryotes.

Because of allelic variation in the reads, many bubbles were created in the De Bruijn graph during the assembly, causing most genes to be partial, with unidentified domains. This situation would make GO analysis of the coding genes less informative. Comparative analysis with coding genes from *Daphnia* with OrthoMCL would create inflated gene families due to broken genes, present on several scaffolds. Additionally, it is not sure if genes present on several scaffolds are paralogs or genes that were broken due to allelic differences.

These analyses will thus only be useful when the number of broken genes in the annotation is decreased, by reduction of the number of scaffolds in the assembly.

4.5.2. Genome annotation

A. franciscana genome annotation characteristics were compared to those of model crustacean *D. pulex* [39], model insect *D. melanogaster* [6,72] and model chelicerate *T. urticae* [72] (Table 4-6).

Table 4-6. Genome annotation characteristics for *A. franciscana* and arthropod models *D. pulex*; *D. melanogaster* and *T. urticae*.

	A. franciscana	*D. pulex*	*D. melanogaster*	*T. urticae*
Number of genes	188,101	31,000	14,861	18,414
Genome size (Mbp)	930	200	180	90
Average gene length (bp)	692	2300	1,506	1,428
Average intron length (bp)	253	170	597	400
Average exon length (bp)	180	210	520	374
Average number of exons per gene	4	7	3	4

Only 1% of the annotated genes could be considered full-length. The fragmentation discussed in section 4.5.1.2. clearly greatly affects the completeness of functional sequence, which affects genome annotation. A very high number of genes were found in *Artemia*, more than 6-fold higher than in the only other annotated crustacean genome of *Daphnia*. When a genome is assembled based solely on NGS reads, the presence of duplicated sequences (genes or segments) and repetitive sequences within introns complicates the complete gene assembly and annotation, often leading to broken genes across multiple sequence scaffolds and sometimes to gene shuffling in the respective scaffolds (i.e. gene exons are in the wrong order, but all on one scaffold) [9]. We have seen several examples of these mechanisms during the *Artemia* genome annotation that might explain why *Artemia* has more and shorter genes with shorter exons in a much larger genome than other model arthropods. *Artemia* genes have a longer average intron length compared to *Daphnia*, but a comparable average number of exons per gene compared to other arthropods.

In order to reduce the number of genes, genome regions with low RNAseq coverage are usually excluded from the genome annotation. This was taken into account in the *Artemia* genome annotation by only using junction data from the RNAseq reads and filtering for at least ten junctions per intron.

Ab initio annotation by comparative analysis with the 38 assembled *Artemia* transcriptomes with the software MAKER2 [81] could be performed to reduce the number of genes even more. In order for this to be efficient, the number of scaffolds in the genome needs to be reduced, to reduce the number of broken genes in the genome.

Most (73%) of the 512 *L. vannamei* nuclear proteins available in the NCBI database, showed homology with the *Artemia* genome. This percentage shows that *Artemia* could be a valid model for a high-value crustacean, such as *L. vannamei*.

4.6. Future research

The *Artemia* genome assembly strategy with CLC Assembly Cell, using paired-end short reads and cre-lox mate-pair reads with scaffolding and gap filling steps proved to result in a useful assembly in spite of the use of short reads from pooled individuals for the assembly of a large genome known for its many repeats [26].

Since the genome and, consequently, the genes are still fragmented in the current assembly, construction of a new assembly based on sequences of an inbred *A. franciscana* and physical or optical mapping (section 5.2.3.1.) would be necessary before GO analysis would be possible. This would provide a deeper understanding of *Artemia franciscana* genes. More details about future research on the *Artemia* genome are provided in Chapter 5.

This *Artemia* genome sequence will allow the study of the influence of extreme environments and the flexibility of extremophiles, uncovering a range of potentially novel genes. Furthermore, the *Artemia* genome will be useful for fine-scale studies of micro evolutionary divergence [95] and for investigation of insect-crustacean branching to better understand the origin and evolution of insects from a crustacean ancestor. Finally, crustacean aquaculture of cash crops, such as *P. monodon* and *L. vannamei*, with often long life cycles and large genomes, would benefit from additional crustacean genomes. As most of the available *L. vannamei* proteins (73%) have homologous sequences in the *Artemia* genome, *Artemia* is a potential model for the high-value crustacean *L. vannamei*.

Chapter 5

Discussion and perspectives

In this chapter, an overview of the results obtained in Chapters 2, 3 and 4 is given, followed by a general discussion focusing on future research and applications for the findings in this work. Finally, concluding remarks are made.

5.1. A genomic toolkit for *Artemia*

In this work, we aimed at developing a genomic toolkit for *Artemia* by:

1) developing **sex-specific AFLP-based genetic linkage maps**, generating sex-linked markers and identifying the sex-determining system present in *Artemia.*

2) discovering **putative genes, involved in primary sex determination in *Artemia*** with a combined strategy of Bulked Segregant Analysis (BSA) and Next-Generation Sequencing (NGS).

3) assembling the *Artemia* genome to its best potential and providing an **annotated draft genome assembly** containing most of the available *Artemia* expression data (ESTs and RNAseq reads).

Accordingly, a genomic *Artemia* toolkit was constructed. Sex-specific AFLP-based genetic maps were constructed based on 433 AFLP markers and 21 male and 22 female linkage groups ($2n = 42$) were identified. Eight sex-linked markers heterozygous in female animals mapped to one locus on a female linkage group, supporting the hypothesis of a WZ/ZZ sex-determining system and showing that primary sex determination is probably directed by one gene (Chapter 2).

To further fine-map the sex locus, BSA was performed (Chapter 3): NGS reads from a male and a female pool were generated, the sequences were assembled *de novo* to an *Artemia* draft genome (Chapter 4) and male and female reads were mapped onto the draft genome. Scaffolds containing genes with SNPs heterozygous in females and homozygous in males were selected and analyzed for gene content. Candidate primary sex-determining genes were identified, including *Cytochrome P450*, *F0F1 ATP synthase subunit beta*, a gene containing a CRAL-TRIO domain, a gene containing an ankyrin repeat, two *Fibronectin* genes, *SEC14* and *Zinc finger C2CH-type* (Chapter 3). Several genes, highly homologous with insect sex-determining genes and crustacean sex-related genes were found in the *Artemia* genome. Except for *Cytochrome P450*, a known candidate sex-determining gene for *M. nipponense*, none of these genes were selected by BSA, indicating that they do not play a role in primary sex determination in *Artemia.*

The 1,310 Mbp *Artemia* draft genome sequence (N50 = 14,784 bp; GC-content = 35%; 176,667 scaffolds) was annotated, predicting 188,101 genes with an average length of 692 bp (Chapter 4). Ninety percent of the known *Artemia* ESTs were present in the *Artemia* draft genome, as well as 92.2% of the RNAseq data in the intermediary genome, revealing that the functional part of the genome under the RNAseq sampling conditions is virtually fully represented in the assembly, although the functional part is still fragmented. As most of the 512 available *L. vannamei* proteins (73%) showed homology with sequences in the *Artemia* genome, *Artemia* could be a potential model for *L. vannamei*.

5.2. General discussion, future research and applications

5.2.1. Linkage maps

Some additional work could still be done on the genetic maps. To determine to which linkage group of the *Artemia* linkage map (Chapter 2) a scaffold from the *Artemia* genome (Chapter 4) belongs, SNPs found in the VarScan analysis (Chapter 3) could be analyzed for segregation in the mapping population by means of a high-density SNP array with, for instance, iSelect HD Custom Genotyping BeadChips (Illumina). This type of analysis would be especially useful for SNPs on scaffolds with putative sex-determining genes (Chapter 3).

Knowledge of the *A. franciscana* sex-determining system (WZ female/ZZ male) will facilitate future evolutionary studies of sex chromosomes in sexually dimorphic (WZ female/ZZ male) and parthenogenetic *Artemia. Artemia* sex-linked markers will enable the study of nauplii sex ratios and their dynamics in laboratory-bred and natural *Artemia* populations. They will also enable the further fine-mapping of the sex-determining locus and the subsequent identification of the primary sex-determining gene(s). Considering the presence of sexual and asexual reproduction strategies, the *Artemia* genus is a promising model system for the study of asexuality, its evolution, and its evolutionary purpose.

5.2.2. Putative primary sex-determining genes

5.2.2.1. Detected by BSA

In the search for sex-determining genes with BSA, several genes (including *Cytochrome P450*, *F0F1 ATP synthase subunit beta*, a gene containing a CRAL-TRIO domain, a gene containing an ankyrin repeat, two *Fibronectin* genes, *Zinc finger C2CH-type* and *SEC14*) with a known link to sex in other animals were detected by BSA as putative genes for sex determination in *Artemia*. The candidate sex-determining genes without a functional description selected by BSA should be analyzed functionally by looking for homologous proteins with Orcae and by producing spline plots, in the same way as for the genes with a functional description (section 3.4.2.). Also, all SNPs in regulatory regions should be investigated to confirm or dismiss the inability of genes to produce sex-specific proteins. Finally, indels found in the *Artemia* genome should be processed and analyzed with the methods applied to SNPs in Chapter 3.

A priority list of BSA-selected putative sex-determining genes to be further investigated would include: *Cytochrome* P450, CRAL-TRIO domain, highly homologous with *SEC14*, proven to influence progeny gender in *Daphnia* and shown to produce sex-specific proteins in *Artemia* and possibly Zinc *finger C2H2*, a current candidate for primary sex determination in *B. mori* (WZ/ZZ). *Cytochrome P450* represents the most valid candidate sex-determining gene; because it produces sex-specific proteins at the secondary protein structure level in *Artemia* and is a candidate sex-determining gene for the crustacean *M. nipponense*, validated by transcriptomic data. There are strong indications that the selected genes are indeed part of the *Artemia* sex-determining pathway and they should be the first genes to be further investigated.

To determine the primary sex-determining gene among the selected candidate genes, further research is needed. Several current methods could be used to confirm or reject the sex-determining character of a candidate gene in *Artemia*.

1) Segregation of selected SNPs with phenotypic sex (5.2.1)

2) Expression of a true primary sex-determining gene should typically start early in development (e.g. embryo) and could persist in adult stages (e.g. gonads). The expression profile of a specific candidate gene could be tested by Reverse-transcription PCR at different developmental stages of *Artemia*. Additionally, qualitative and differential expression of the putative genes could be tested by performing RNAseq of adult males and females and comparing the transcript sequence(s) in both sexes.

3) A true sex-determining gene among selected candidates could be identified, for instance, by interfering with the expression of putative sex-determining genes by means of antisense oligonucleotides [148] in juvenile *Artemia* at the appropriate time points (based on expression profiles) with subsequent observation of the effect on phenotypic sex (ratios).

4) Past transformation methods of *Artemia* include ballistic introduction of transient luciferase [68] and introduction of green fluorescent protein and growth hormone by electroporation [30]. CRISPR-Cas genome editing could be used to knock out candidate sex-determining genes in *Artemia*. CRISPR-Cas can knock out/ knock in genes or make allelic mutants, target multiple sites in a genome, and has already been used on many organisms. It only requires design and synthesis of guide RNA [180] and injection into the zygote, so a decapsulated *Artemia* cyst could be used.

5) The biochemical function of a gene, often predicted by annotation, can be a good indication for the role a gene could play in the sex determination cascade, provided that this cascade or parts of it are known in the studied organism or closely-related organisms. Moreover, many sex-determining genes have been found to be transcription factors (such as *SRY* and *DMRT1*) [148].

Once the sex-determining gene is identified, characterization of target genes regulated by the sex-determining gene would be the next interesting route for future research. Also, based on sequence homology with *Artemia*, sex-determining genes might be identified in commercially valuable crustaceans, enabling PCR-based allele-specific assay development in the framework of the development of mono-sex cultures in shrimp [216].

5.2.2.2. Arthropod sex-determining gene homologs, not detected by BSA

Genes, not selected by BSA, but homologous to known insect sex-determining genes (*dsx*, *fru*, *SRY*, *sxl*, *tra-1*, *tra-2*, *fem1*, *runt*, *dpn*, *da*, *emc*, *gro* and *sf*) [132] are present in the *Artemia* genome comprising most of the genes in the *Drosophila* sex determination cascade. *Dsx* and *tra* have been shown to play a role in sex determination for certain crustaceans: in *Daphnia*, *DapmaDsx1* is responsible for male trait development during ESD [12,76] and in *F. chinensis*, the expression of *FcTra-2c*is is significantly higher in juvenile females than in males, with the highest expression levels in ovary tissues [14]. It would be interesting to look for non-synonymous SNPs in these genes, because their respective genome fragments might not have contained enough SNPs (a minimum of five SNPs per scaffold was demanded) to be picked up by the BSA analysis, whereas, in principle, one SNP can be enough for the production of a different protein. The most interesting genes to study would be *tra*, *fem1*, *sxl* and *dsx*, because these genes are responsible for primary sex determination in arthropods.

5.2.3. The genome

5.2.3.1. Fragmentation

The *Artemia* genome assembly is larger than the estimated genome size, hinting at redundancy and the assembly with over 170 K scaffolds remains fragmented, greatly affecting subsequent genome annotation. Redundancy is due to fragmentation, in turn due to allelic variation in sequenced pools, unresolved haplotypes because of the use of diploid organisms and the high content of repetitive elements, combined with the fact that NGS reads cannot span genomic repeats longer than the read insert sizes of this study (200 bp and 500 bp for PE; 3000 bp for MP). Once the coverage plateau is reached, additional paired-end sequencing will never resolve the assembly [122]. As for any genome project, post-assembly genome finishing steps are still needed, such as large-insert MP sequencing, pseudo-Sanger sequencing, single-molecule sequencing and physical or optical mapping.

Fragmentation due to differences between homologous chromosomes due to diploidy could be reduced by sequencing DNA from an *A. franciscana* inbred line, removing variation due to pooling of individuals from the assembly. Fragmentation due to repeats could be reduced by applying "long read sequencing" or "long insert mate-pair sequencing". Producing a new genome assembly following these criteria would ultimately lead to a less fragmented (and thus less redundant) assembly, containing less broken genes.

Nambu *et al*. (2007) have developed an inbred *A. franciscana* strain by full-sib crossing for sixty generations [147]. One or more female inbred individuals (including brood pouch, free of embryos in order to extract enough DNA for several libraries) would be sequenced. Using pools of full inbreds does not have an influence on allelic variation. Sequencing females (WZ) would provide more information about *Artemia* sex chromosomes than males (ZZ).

As mentioned in Chapter 1, methods exist to elongate reads from NGS platforms, termed "pseudo-Sanger sequencing". This method starts with short reads from a series of PE libraries of stepwise decreasing insert sizes (maximal insert size 800 bp), that are computationally transformed into near error-free pseudo-Sanger sequences with the length of the largest insert size (800 bp) [176]. Pseudo-Sanger reads have three advantages over normal, untransformed PE reads: better gap filling, error correction and heterozygote tolerance [176]. When tested for *de novo* assembly on a non-isogenic strain of *D. melanogaster*, pseudo-Sanger sequencing yielded a N50 contig of 190 kb, a five-fold improvement over the existing *de novo* short read assembly methods and a three-fold advantage over the assembly of reads from 454 sequencing [176]. Sanger sequencing itself is not taken into consideration, due to its prohibitive cost.

Combined MP libraries with insert sizes that match the distributions of repetitive elements improve scaffolding and can contribute to finalizing draft genomes [209]. In *Rattus norvegicus*, the combination of medium (8-15 Kb) and large insert libraries (20-25 Kb) resulted in a 3-fold increase in N50 in scaffolding processes [209]. The question is how long the repeats impeding the *Artemia* genome assembly are. Because precisely these repeats can be the reason for assembly failures with NGS reads, this question remains difficult to answer. When closely-related sequenced organisms are considered, such as *D. pulex*, TELSAT1 repeats form sub-telomeric clusters over 40 kb in length [39], indicating that assembly problems due to repeats can be greatly improved, but never completely resolved with NGS or even Sanger sequencing in genomes containing long repeats.

Emerging technologies, such as single-molecule real-time (SMRT) sequencing [185] may eventually help to overcome many of the genome assembly challenges put forward in this study, because they produce long reads, with the added benefits of a low GC bias and no amplification bias. Genomes solely based on this technique have so far mostly been published for bacteria and not for large, complex organisms, probably due to the prohibitive costs for large genomes [185]. Recent unpublished results for genome assembly of the higher plant model *Arabidopsis thaliana* show however that large genomes can be assembled, based on SMRT sequencing only [3]. Single-molecule technologies are more prone to sequencing errors than NGS, but because the errors are random (and not systematic, as in NGS), they can be neutralized by using single-molecule reads as the `back bone` of the assembly, and using NGS reads for base correction in this `back bone` (an approach called hybrid assembly [47]). Application of methods solely based on current NGS technology remain the most straight-forward choice, because such platforms mature fast and are very broadly available and affordable. Also, software for NGS data analysis such as for genome assembly is readily available, whereas all-in-one assembly software for hybrid assemblies has not been developed yet.

An additional manner to address issues with redundancy, fragmentation and repeats, once longer scaffold lengths have been reached using the methods described above, is optical mapping with BionanoGenomics technology [107]. This technology uses enzymes to create sequence-specific nicks that are subsequently labeled by a fluorescent nucleotide analog. The 100-Kb-long nick-labeled DNA fragments are stained and loaded onto a nanofluidic chip by an electric field. The DNA is linearized by confinement in a nanochannel array, resulting in uniform linearization and allowing precise and accurate measurement of the distance between the nick labels on the DNA molecules, forming a signature pattern. Signature patterns of all DNA fragments are imaged and aligned, creating a consensus optical map. Continuing efforts to assemble the *T. urticae* genome [72], by means of BionanoGenomics optical mapping technology with one nicking enzyme have led to more than a 100-fold reduction of the amount of fragments in the genome assembly (Stephane Rombauts, personal communication). Optimally, this technology would enable production of chromosome-length scaffolds, but two factors need to be taken into account: (1) the required length of the DNA molecules (100 Kb) for optical mapping, that is not straight-forward to obtain with current DNA extraction methods and (2) the large amount of short scaffolds still present in the current *Artemia* genome assembly that will probably contain no more than one nick and, thus, not be mapped.

BionanoGenomics optical mapping can however be applied with two nicking enzymes. A two-color labeling strategy resulted in an average information density of one label per 4.8 Kb in the large and complex genome of *Aegilops tauschii*, anchoring unplaced sequence contigs, validating the initial draft assembly, resolving instances of misassembly, some involving contigs <2 kb long and improving the assembly dramatically to a coverage of 95% of the consensus genome map [76]. Supposing an average marker density of 4.8 KB, an optical map of the *Artemia* genome containing approximately 300,000 markers could be created, resolving most fragments except scaffolds shorter than 5 Kb.

Although not described yet in the literature, an additional two-color labeled run is possible, with one nicking enzyme in common with the previous run and one new nicking enzyme. This corresponds to one single run with three nicking enzymes at the same time, which is currently not possible due to image resolution issues. Such an assay would improve the marker density even more, allowing more scaffolds shorter than 5 kb to be resolved. In the end, to assemble repeats, connect oriented scaffolds and correct misassemblies, no better strategies are available on the market than optical mapping with BionanoGenomics, because it is fast (one day per run), examines long genome stretches (minimum 100 Kb) without breaking its original structure and is affordable (estimated cost of an *Artemia* optical map made with three chips: € 3,000).

Improving the *Artemia* assembly will reduce the number of scaffolds, the gene fragmentation and enhance gene structure, thus greatly improving the subsequent annotation and biological significance of the assembly.

5.2.3.2. Differential expression

After improving of the *Artemia* assembly, the annotation will be further improved by manual curation by all partners involved in the *Artemia* genome project. Each of the partners will elaborate a specific research area, by using the *Artemia* genome and the results of differential expression analysis with RNAseq data from *Artemia* under different conditions and life cycle stages (Chapter 4). Research groups involved in this project will study different metabolic states, such as diapause and quiescence and study the effects of different types of stress, such as desiccation and anoxia. Also, research groups focusing on cell cycle research and, more specifically, on cell cycle arrest in diapaused *Artemia*, will benefit from the generated RNAseq data. Finally, laboratories performing crustacean immunology research are interested in the immunity-related genes expressed by *Artemia*. Typical genes, related to the *Artemia* life cycle and stress resistance will be discovered during this collaboration.

The already discovered *Artemia* genes such as *APH-1* (osmoregulation), *RSK* (cell cycle arrest termination in cysts) and *LEA* (development of bio-stable dried cells) described in Chapter 1, show the potential discoveries that are still to be made in the *Artemia* genome.

5.2.3.3. Comparative genomics, pathway mapping and QTL analysis

After improving the *Artemia* assembly and annotation, comparative genomics studies with the software package OrthoMCL to compare the *Artemia* genome with all sequenced arthropod genomes, such as *Daphnia* and *Drosophila* will provide a more complete image of the uniqueness of *Artemia* within the arthropods. Comparative studies of crustacean and insect gene families will enable the study of crustacean evolution and origins. *Artemia* is particularly well-suited for this purpose because it is close to the insect-crustacean branching in the phylogenetic tree (Figure 1-1). Also, some homologies with worms, such as *C. elegans* have been observed in the annotated *Artemia* genome, so it could be interesting to include *C. elegans* proteins in *Artemia* comparative genomics studies. In the future, when more crustacean genomes will be sequenced, comparative genomics will reveal common gene families between *Artemia* and other crustaceans as well. Only then, it will be possible to comprehensively evaluate the value of *Artemia* as a model crustacean, specifically for high-value cultured crustaceans, such as *P. monodon* and *L. vannamei*. This kind of evaluation is currently going on for *Drosophila*, which, after completion of many insect genome projects, may turn out to be a less representative insect model than previously expected [127].

Another interesting tool to apply on the protein data contained in an improved *Artemia* genome annotation, and on the RNAseq data, is KEGG PATHWAY mapping. KEGG PATHWAY maps large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretation of higher-level systemic functions (http://www.genome.jp/kegg/pathway.html).

The availability of genome sequences enables custom-made molecular markers that are tightly linked to target loci. The tight relationship between linkage map and genome sequence could further help to understand which genes underlie specific quantitative trait loci. Genome sequences also offer the opportunity to perform genome-wide association studies that are powerful tools for high-resolution mapping of complex quantitative traits.

5.3. Conclusion

In this study, we showed that genetic linkage mapping remains an efficient tool to support genetic studies in the genomics era. Biology can make significant progress through integrative studies merging genomics, phenotyping, and classical genetics.

In an effort to put *Artemia* forward as a new genomic model for crustaceans, several steps have been taken in this study. Sex-specific high-density AFLP-based genetic linkage maps containing sex-linked markers demonstrated that *Artemia* has a WZ/ZZ sex-determining system, with primary sex mapping to a single locus. Several candidate genes connected with primary sex determination in arthropods were identified (*Zinc finger, C2H2*; *CRAL-TRIO* and *Cytochrome P450*). Further testing of these candidate genes should narrow down this selection towards one primary sex-determining gene for *Artemia*. An *Artemia* genome sequence was assembled *de novo* and annotated.

Although the functional part of the genome under the RNAseq sampling conditions is virtually fully represented in the assembly thus making it useful for qualitative research, genome finishing strategies, such as large-insert MP and pseudo-Sanger or single-molecule sequencing of an inbred, followed by optical mapping will still be necessary to complete the genome project.

The further development of genomic resources for *Artemia* will add a completely new dimension to *Artemia* research. As *Artemia* is considered a potential crustacean model species, with protein homology with the high-value crustacean *L. vannamei,* increasing knowledge about *Artemia* genetics and genomics in general, and sex-related genetics in particular, are expected to be valuable to crustacean aquaculture, presently lacking molecular breeding strategies despite their high contribution to the total aquaculture production value. The increasing sequencing efforts applied to crustaceans through consortia such as the Global Invertebrate Genomics Alliance [70] and the 5,000 Insect Genome Project [1], as well as the finished *Artemia* genome will broaden crustacean genomics in general and enable *Artemia* genomics to move forward as more crustacean genomes become available.

References

1. i5k nomination summary

http://arthropodgenomes.org/wiki/i5K_nominations

2. Sequenced arthropod genomes

http://arthropodgenomes.org/wiki/Sequenced_genomes

3. New data release: *Arabidopsis* assembly offers glimpse of *de novo* SMRT sequencing for larger genomes. (2013).

http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html

4. Abatzopoulos TJ, Beardmore JA, Clegg JS, Sorgeloos P (2002). *Artemia*: basic and applied biology. Dordrecht, Kluwer Academic Publishers. 304 p.

5. Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012). GenomeView: a next-generation genome browser. Nucleic Acids Research, 40:2, e12.

6. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P,

Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000). The genome sequence of *Drosophila melanogaster*. Science, 287:5461, 2185-2195.

7. Agh N, Abatzopoulos TJ, Kappas I, Van Stappen G, Rouhani SMR, Sorgeloos P (2007). Coexistence of sexual and parthenogenetic *Artemia* populations in Lake Urmia and neighbouring lagoons. International Review of Hydrobiology, 92:1, 48-60.
8. Alcivar-Warren A (2012). The plasticity of the shrimp genome -sex, retrotransposons, ribosomal RNAs, growth performance and disease susceptibility. Aquaculture America, International Marine Shrimp Environmental Genomics Initiative (IMSEGI) – Monitoring ecosystem, animal and public health. Las Vegas, Nevada.
9. Alkan C, Sajjadian S, Eichler EE (2011). Limitations of next-generation genome sequence assembly. Nature Methods, 8:1, 61-65.

10. Allison DB, Gadbury. G.L., Heo M, Fernandez JR, Lee C-K, Prolla TA, Weindruch R (2002). A mixture model approach for the analysis of microarray gene expression data. Computational Statistics & Data Analysis, 39:-, 1-16.

11. Andrenacci D, Le Bras S, Grimaldi MR, Rugarli E, Graziani F (2004). Embryonic expression pattern of the *Drosophila* Kallmann syndrome gene *kal-1*. Gene Expression Patterns, 5:1, 67-73.

12. Asem A, Rastegar-Pouyani N, De Los Ríos-Escalante P (2010). The genus *Artemia* Leach, 1819 (Crustacea: Branchiopoda). I. True and false taxonomical descriptions. Latin American Journal of Aquatic Research, 38:3, 501-506.

13. Azzouna A, Greve P, Martin G (2004). Sexual differentiation traits in functional males with female genital apertures (male symbol fga) in the woodlice *Armadillidium vulgare* Latr. (Isopoda, Crustacea). General and Comparative Endocrinology, 138:1, 42-49.

14. Badaracco G, Bellorini M, Landsberger N (1995). Phylogenetic study of bisexual *Artemia* using random amplified polymorphic DNA. Journal of Molecular Evolution, 41:2, 150-154.

15. Baldwin WS, Marko PB, Nelson DR (2009). The cytochrome P450 (CYP) gene superfamily in *Daphnia pulex*. BMC Genomics, 10:169.

16. Baxevanis AD, Kappas I, Abatzopoulos TJ (2006). Molecular phylogenetics and asexuality in the brine shrimp *Artemia*. Molecular Phylogenetics and Evolution, 40:3, 724-738.

17. Benzie JAH (2009). Use and exchange of genetic resources of penaeid shrimps for food and aquaculture. Reviews in Aquaculture, 1:3-4, 232-250.

18. Bergero R, Charlesworth D (2009). The evolution of restricted recombination in sex chromosomes. Trends in Ecology & Evolution, 24:2, 94-102.

19. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics, 27:4, 578-579.

20. Boetzer M, Pirovano W (2012). Toward almost closed genomes with GapFiller. Genome Biology, 13:6, R56.

21. Bossier P, Xiaomei W, Catania F, Dooms S, Van Stappen G, Naessens E, Sorgeloos P (2004). An RFLP database for authentication of commercial cyst samples of the brine shrimp *Artemia* spp. (International Study on *Artemia* LXX). Aquaculture, 231:1-4, 93-112.

22. Bowen ST (1965). The genetics of *Artemia salina*. V. Crossing over between X and Y chromosomes. Genetics, 52:3, 695-710.

23. Browne RAB, S.T. (1991). Taxonomy and population genetics of *Artemia*. In. *Artemia* biology. Boca Raton, FL.: CRC Press. pp. 221-235.

24. Browne RAW, G. (2000). Combined effects of salinity and temperature on survival and reproduction of five species of *Artemia*. Journal of Experimental Marine Biology and Ecology, 244:1, 29-44.

25. Camargo WN, Bossier P, Sorgeloos P, Sun Y (2002). Preliminary genetic data on some Caribbean *Artemia franciscana* strains based on RAPD's. Hydrobiologia, 468:1-3, 245-249.

26. Carettoni D, Landsberger N, Zagni E, Benfante R, Badaracco G (1994). Topoisomerase-I action on the heterochromatic DNA from the brine shrimp *Artemia franciscana*: Studies *in vivo* and *in vitro*. Biochemical Journal, 299:3, 623-629.

27. Carmichael SN, Bekaert M, Taggart JB, Christie HR, Bassett DI, Bron JE, Skuce PJ, Gharbi K, Skern-Mauritzen R, Sturm A (2013). Identification of a Sex-Linked SNP Marker in the Salmon Louse (*Lepeophtheirus salmonis*) Using RAD Sequencing. PLoS One, 8:10, e77832.

28. Carter JM, Baker SC, Pink R, Carter DR, Collins A, Tomlin J, Gibbs M, Breuker CJ (2013). Unscrambling butterfly oogenesis. BMC Genomics, 14:-, 283.

29. The brine shrimp life cycle. (2002) Genetic science learning center.

http://learn.genetics.utah.edu/content/gsl/artemia/

30. Chang SH, Lee BC, Chen YD, Lee YC, Tsai HJ (2011). Development of transgenic zooplankton *Artemia* as a bioreactor to produce exogenous protein. Transgenic Research, 20:5, 1099-1111.

31. Chavez M, Landry C, Loret S, Muller M, Figueroa J, Peers B, Rentier-Delrue F, Rousseau GG, Krauskopf M, Martial JA (1999). APH-1, a POU homeobox gene expressed in the salt gland of the crustacean *Artemia franciscana*. Mechanisms of Development, 87:1-2, 207-212.

32. Chistiakov DA, Hellemans B, Volckaert FAM (2006). Microsatellites and their genomic distribution, evolution, function and applications: A review with special reference to fish genetics. Aquaculture Europe, 255: :1-4, 1-29.

33. Cho S, Huang ZY, Zhang J (2007). Sex-specific splicing of the honeybee *doublesex* gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway. Genetics, 177:3, 1733-1741.

34. Claesen J, Clement L, Shkedy Z, Foulquie-Moreno MR, Burzykowski T (2013). Simultaneous Mapping of Multiple Gene Loci with Pooled Segregants. PLoS One, 8:2, e55133.

35. CLC bio (2012). White paper on de novo assembly in CLC Assembly Cell 4.0.

36. Clegg JS (1997). Embryos of *Artemia franciscana* survive four years of continuous anoxia: the case for complete metabolic rate depression. The Journal of Experimental Biology, 200:3, 467-475.

37. Clegg JS (2005). Desiccation tolerance in encysted embryos of the animal extremophile, *Artemia*. Integrative and Comparative Biology, 45:5, 715-724.

38. Clegg JS (2007). Protein stability in *Artemia* embryos during prolonged anoxia. Biological Bulletin, 212:1, 74-81.

39. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR,

Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011). The ecoresponsive genome of *Daphnia pulex*. Science, 331:6017, 555-561.

40. Copf T, Schroder R, Averof M (2004). Ancestral role of caudal genes in axis elongation and segmentation. Proceedings of the National Academy of Sciences of the United States of America, 101:51, 17711-17715.
41. Coutteau P, Brendonck L, Lavens P, Sorgeloos P (1992). The Use of manipulated baker's yeast as an algal substitute for the laboratory culture of Anostraca. Hydrobiologia, 234:1, 25-32.
42. Crease TJ (1999). The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). Gene, 233:1-2, 89-99.
43. Cristescu MEA, Colbourne JK, Radivojac J, Lynch M (2006). A micro satellite-based genetic linkage map of the waterflea, *Daphnia pulex*: On the prospect of crustacean genomics. Genomics, 88:4, 415-430.
44. Cruces J, Wonenburger MLG, Díaz-Guerra M, Sebastián J, Renart J (1986). Satellite DNA in the crustacean *Artemia*. Gene, 44:2-3, 341-345.
45. Dai JQ, Zhu XJ, Liu FQ, Xiang JH, Nagasawa H, Yang WJ (2008). Involvement of p90 ribosomal S6 kinase in termination of cell cycle arrest during development of *Artemia*-encysted embryos. The Journal of Biological Chemistry, 283:3, 1705-1712.
46. Degroeve S, Saeys Y, De Baets B, Rouze P, Van de Peer Y (2005). SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics, 21:8, 1332-1338.
47. Deshpande V, Fung EDK, Pham S, Bafna V (2013). Cerulean: A hybrid assembly using high throughput short and long reads. 13th Workshop on Algorithms in Bioinformatics. France.
48. Desjardins P, Conklin D (2010). NanoDrop microvolume quantitation of nucleic acids. Journal of Visualized Experiments, 45.

49. Diz AP, Dudley E, MacDonald BW, Pina B, Kenchington EL, Zouros E, Skibinski DO (2009). Genetic variation underlying protein expression in eggs of the marine mussel *Mytilus edulis*. Molecular & Cellular Proteomics, 8:1, 132-144.

50. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research, 36:16, e105.

51. Doležel J, Bartoš J (2005). Plant DNA flow cytometry and estimation of nuclear genome size. Annals of Botany, 95:1, 99-110.

52. Du ZQ, Ciobanu DC, Onteru SK, Gorbach D, Mileham AJ, Jaramillo G, Rothschild MF (2010). A gene-based SNP linkage map for pacific white shrimp, *Litopenaeus vannamei*. Animal Genetics, 41:3, 286-294.

53. Escalante R, Garcia-Saez A, Ortega MA, Sastre L (1994). Gene expression after resumption of development of *Artemia franciscana* cryptobiotic embryos. Biochemistry and Cell Biology, 72:3-4, 78-83.

54. Eto K, Sonoda Y, Abe S (2011). The kinase DYRKIA regulates pre-mRNA splicing in spermatogonia and proliferation of spermatogonia and Sertoli cells by phosphorylating a spliceosomal component, SAP155, in postnatal murine testes. Molecular and Cellular Biochemistry, 355:1-2, 217-222.

55. FAO (2010). The State of World Fisheries and Aquaculture. World review of fisheries and aquaculture. Rome. 89 p.

56. FAO (2010). FAO - Fisheries and Aquaculture Information and Statistics Service.

57. Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Van de Peer Y, Rouze P, Schiex T (2008). Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics, 3:2, 87-97.

58. Foley BR, Rose CG, Rundle DE, Leong W, Moy GW, Burton RS, Edmands S (2011). A gene-based SNP resource and linkage map for the copepod *Tigriopus californicus*. BMC Genomics, 12.

59. Ford AT (2008). Can you feminise a crustacean? Aquatic Toxicology, 88:4, 316-321.

60. The salmon louse genome sequencing project: part 1. (2012) Fiskeri- og havbruksnaeringens forskningsfond.

http://www.fhf.no/prosjektdetaljer/?projectNumber=900400

61. Freitas MCR, António JMS, Ziolli RL, Yoshida MI, Rey NA, Diniz R (2011). Synthesis and structural characterization of a zinc(II) complex of the mycobactericidal drug isoniazid – Toxicity against *Artemia salina*. Polyhedron, 30:11, 1922-1926.

62. Fujii T, Shimada T (2007). Sex determination in the silkworm, *Bombyx mori*: a female determinant on the W chromosome and the sex-determining gene cascade. Seminars in Cell & Developmental Biology, 18:3, 379-388.

63. Gailey DA, Billeter JC, Liu JH, Bauzon F, Allendorfer JB, Goodwin SF (2006). Functional conservation of the *fruitless* male sex-determination gene across 250 Myr of insect evolution. Molecular Biology and Evolution, 23:3, 633-643.

64. Gajardo G BJ (2001). Coadaptation: lessons from the brine shrimp *Artemia*, the aquatic *Drosophila* (crustacea, Anostraca). Revista Chilena de Historia Natural, -:74, 65-72.

65. Gajardo GM, Sorgeloos P, Beardmore JA (2006). Inland hypersaline lakes and the brine shrimp *Artemia* as simple models for biodiversity analysis at the population level. Saline Systems, 2:14, -.

66. Gajardo GM, Beardmore JA (2012). The brine shrimp *Artemia*: adapted to critical life conditions. Frontiers in Physiology, 3:185.

67. Garesse R, Carrodeguas JA, Santiago J, Perez ML, Marco R, Vallejo CG (1997). *Artemia* mitochondrial genome: molecular biology and evolutive considerations. Comparative Biochemistry and Physiology Part B, Biochemistry & Molecular Biology, 117:3, 357-366.

68. Gendreau SL, V.; Cadoreta, J.P.; Mialhea, E. (1995). Transient expression of a luciferase reporter gene after ballistic introduction into *Artemia franciscana* (Crustacea) embryos. Aquaculture Europe, 133:3-4, 199-205.

69. Genet C, Dehais P, Palti Y, Gao G, Gavory F, Wincker P, Quillet E, Boussaha M (2011). Analysis of BAC-end sequences in rainbow trout: Content characterization and assessment of synteny between trout and other fish genomes. BMC Genomics, 12:314.

70. Global Invertebrate Genomics Alliance. (2014) GIGA.

http://giga.nova.edu/

71. Gomez-Uchida D, Weetman D, Hauser L, Galleguillos R, Retamal M (2003). Allozyme and AFLP analyses of genetic population structure in the hairy edible crab *Cancer setosus* from the Chilean coast. Journal of Crustacean Biology, 23:2, 486-494.

72. Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, Grbic V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, Hernandez-Crespo P, Diaz I, Martinez M, Navajas M, Sucena E, Magalhaes S, Nagy L, Pace RM, Djuranovic S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi SV, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y (2011). The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature, 479:7374, 487-492.

73. Animal Genome Size Database. (2005) Gregory TR.

http://www.genomesize.com

74. Griffiths AJF, Wessler SR, Lewontin RC, Carroll SB (2008). "Mapping eukaryote chromosomes by recombination" in Introduction to Genetic Analysis New York, W.H. Freeman and Company.

75. Haldar S, Chatterjee S, Sugimoto N, Das S, Chowdhury N, Hinenoya A, Asakura M, Yamasaki S (2011). Identification of *Vibrio campbellii* isolated from diseased farm-shrimps from south India and establishment of its pathogenic potential in an *Artemia* model. Microbiology, 157:1, 179-188.

76. Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J, Luo MC, Gu Y, Xiao M (2013). Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. PLoS One, 8:2, e55864.

77. Hauser F, Cazzamali G, Williamson M, Park Y, Li B, Tanaka Y, Predel R, Neupert S, Schachtner J, Verleyen P, Grimmelikhuijzen CJ (2008). A genome-wide inventory of neurohormone GPCRs in the red flour beetle *Tribolium castaneum*. Frontiers in Neuroendocrinology, 29:1, 142-165.

78. Hengherr S, Schill RO, Clegg JS (2011). Mechanisms associated with cellular desiccation tolerance in the animal extremophile *Artemia*. Physiological and Biochemical Zoology, 84:3, 249-257.

79. Heule C, Salzburger W, Bohne A (2014). Genetics of sexual development: an evolutionary playground for fish. Genetics, 196:3, 579-591.

80. Hodgson R (1999). CTAB method for the isolation of total nucleic acid (TNA) from shrimp tissue. Workshop on “Molecular diagnostics for shrimp viruses in the Asian region”. Salaya.

81. Holt C, Yandell M (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics, 12:491.

82. Hsu W-J, Chichester CO, Davies BH (1970). The metabolism of β-carotene and other carotenoids in the brine shrimp, *Artemia salina* L. (Crustacea: Branchiopoda). Comparative Biochemistry and Physiology, 32:1, 69-79.

83. Huang W, Richards S, Carbone MA, Zhu D, Anholt RR, Ayroles JF, Duncan L, Jordan KW, Lawrence F, Magwire MM, Warner CB, Blankenburg K, Han Y, Javaid M, Jayaseelan J, Jhangiani SN, Muzny D, Ongeri F, Perales L, Wu YQ, Zhang Y, Zou X, Stone EA, Gibbs RA, Mackay TF (2012). Epistasis dominates the genetic architecture of *Drosophila* quantitative traits. Proceedings of the National Academy of Sciences of the United States of America, 109:39, 15553-15559.

84. Innan H, Terauchi R, Kahl G, Tajima F (1999). A method for estimating nucleotide diversity from AFLP data. Genetics, 151:3, 1157-1164.

85. The salmon louse genome project. Institute of Marine Research, Bergen.

http://sealouse.imr.no/

86. VSN International (2011). GenStat for Windows 14th Edition. Hemel Hempstead, UK.

87. Jager S, Schwartz HT, Horvitz HR, Conradt B (2004). The *Caenorhabditis elegans* F-box protein SEL-10 promotes female development and may target FEM-1 and FEM-3 for degradation by the proteasome. Proceedings of the National Academy of Sciences of the United States of America, 101:34, 12549-12554.

88. Jeffery NW (2012). The first genome size estimates for six species of krill (Malacostraca, Euphausiidae): large genomes at the north and south poles. Polar Biology, 35:6, 959-962.

89. Jiang G, Xu X, Jing Y, Wang R, Fan T (2011). Comparative studies on sorting cells from *Artemia sinica* at different developmental stages for *in vitro* cell culture. In Vitro Cell Developmental Biology Animal, 47:5-6, 341-345.

90. Jin S, Fu H, Zhou Q, Sun S, Jiang S, Xiong Y, Gong Y, Qiao H, Zhang W (2013). Transcriptome analysis of androgenic gland for discovery of novel genes from the oriental river prawn, *Macrobrachium nipponense*, Using Illumina Hiseq 2000. PLoS One, 8:10, e76840.

91. Juchault P, Rigaud T (1995). Evidence for female heterogamety in two terrestrial crustaceans and the problem of sex chromosome evolution in isopods. Heredity, 75:466-471.

92. Jung H, Lyons RE, Dinh H, Hurwood DA, McWilliam S, Mather PB (2011). Transcriptomics of a giant freshwater prawn (Macrobrachium rosenbergii): de novo assembly, annotation and marker discovery. PLoS One, 6:12, e27938.

93. Kaiser VB, Bachtrog D (2010). Evolution of sex chromosomes in insects. Annual Review of Genetics, 44 -, 91-112.

94. Kamiya T, Kai W, Tasumi S, Oka A, Matsunaga T, Mizuno N, Fujita M, Suetake H, Suzuki S, Hosoya S, Tohari S, Brenner S, Miyadai T, Venkatesh B, Suzuki Y, Kikuchi K (2012). A trans-species missense SNP in *Amhr2* is associated with sex determination in the tiger pufferfish, *Takifugu rubripes* (Fugu). Plos Genetics, 8:7, e1002798.

95. Kappas I, Abatzopoulos TJ, Van Hoa N, Sorgeloos P, Beardmore JA (2004). Genetic and reproductive differentiation of *Artemia franciscana* in a new environment. Marine Biology, 146:1, 103-117.

96. Kato Y, Kobayashi K, Oda S, Tatarazako N, Watanabe H, Iguchi T (2010). Sequence divergence and expression of a *transformer* gene in the branchiopod crustacean, *Daphnia magna*. Genomics, 95:3, 160-165.

97. Kato Y, Kobayashi K, Watanabe H, Iguchi T (2011). Environmental sex determination in the branchiopod crustacean *Daphnia magna*: deep conservation of a *doublesex* gene in the sex-determining pathway. Plos Genetics, 7:3, e1001345.

98. Kayim M, Ates M, Elekon HA (2010). The effects of different feeds under the same salinity conditions on the growth and survival rate of *Artemia*. Journal of Animal and Veterinary Advances, 9:8, 1223-1226.

99. Kelley DR, Schatz MC, Salzberg SL (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biology, 11:11, R116.

100. Kelly WG, Aramayo R (2007). Meiotic silencing and the epigenetics of sex. Chromosome Research, 15:5, 633-651.

101. Khamnamtong B, Thumrungtanakit S, Klinbunga S, Aoki T, Hirono I, Menasveta P (2006). Identification of sex-specific expression markers in the giant tiger shrimp (*Penaeus monodon*). Journal of Biochemistry and Molecular Biology, 39:1, 37-45.

102. Khmeleva NN, Iurkevich GN (1968). Energy metabolism in *Artemia salina* (L.) and other crustacea. Doklady Akademii nauk SSSR, 183:4, 978-981.

103. Killian DJ, Harvey E, Johnson P, Otori M, Mitani S, Xue D (2008). SKR-1, a homolog of Skp1 and a member of the SCF(SEL-10) complex, regulates sex-determination and LIN-12/Notch signaling in *C. elegans*. Deveopmental Biology, 322:2, 322-331.

104. King AM, MacRae TH (2012). The Small Heat Shock Protein p26 Aids Development of Encysting Artemia Embryos, Prevents Spontaneous Diapause Termination and Protects against Stress. PLoS One, 7:8.

105. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25:17, 2283-2285.

106. Kopp A (2012). *Dmrt* genes in the development and evolution of sexual dimorphism. Trends in Genetics, 28:4, 175-184.

107. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M, Kwok PY (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature Biotechnology, 30:8, 771-776.

108. Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K (2011). Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. Genetics, 188:2, 239-246.

109. Lavens P, Sorgeloos P (1996). Manual on the production and use of live food for aquaculture. FAO Fisheries Technical Paper 361; FAO, editor. Rome. 295 p.

110. Leelatanawit R, Sittikankeaw K, Yocawibun P, Klinbunga S, Roytrakul S, Aoki T, Hirono I, Menasveta P (2009). Identification, characterization and expression of sex-related genes in testes of the giant tiger shrimp *Penaeus monodon*. Comparative Biochemistry and Physiology Part A, Molecular & Integrative Physiology, 152:1, 66-76.

111. Leger P, Bengtson DA, Simpson KL, Sorgeloos P (1986). The use and nutritional value of *Artemia* as a food source. Oceanography and Marine Biology, 24:-, 521-623.

112. Leggett RM, Ramirez-Gonzalez RH, Verweij W, Kawashima CG, Iqbal Z, Jones JD, Caccamo M, Maclean D (2013). Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. PLoS One, 8:3, e60058.

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

114. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25:16, 2078-2079.

115. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J (2010). The sequence and *de novo* assembly of the giant panda genome. Nature, 463:7279, 311-317.

116. Li S, Chakraborty N, Borcar A, Menze MA, Toner M, Hand SC (2012). Late embryogenesis abundant proteins protect human hepatoma cells during acute desiccation. Proceedings of the National Academy of Sciences of the United States of America, 109:51, 20859-20864.

117. Li S, Li F, Wen R, Xiang J (2012). Identification and characterization of the sex-determiner *transformer-2* homologue in Chinese shrimp, *Fenneropenaeus chinensis*. Sexual Development, 6:5, 267-278.

118. Li Y, Byrne K, Miggiano E, Whan V, Moore S, Keys S, Crocos P, Preston N, Lehnert S (2003). Genetic mapping of the kuruma prawn *Penaeus japonicus* using AFLP markers. Aquaculture, 219:1-4, 143-156.

119. Li YT, Dierens L, Byrne K, Miggiano E, Lehnert S, Preston N, Lyons R (2006). QTL detection of production traits for the Kuruma prawn *Penaeus japonicus* (Bate) using AFLP markers. Aquaculture, 258:1-4, 198-210.

120. Li ZX, Li J, Wang QY, He YY, Liu P (2006). AFLP-based genetic linkage map of marine shrimp *Penaeus* (*Fenneropenaeus*) *chinensis*. Aquaculture, 261:2, 463-472.

121. Libertini A, Trisolini R, Rampin M (2008). Chromosome number, karyotype morphology, heterochromatin distribution and nuclear DNA content of some talitroidean amphipods (Crustacea : Gammaridea). European Journal of Entomology, 105:1, 53-58.

122. Lin Y, Li J, Shen H, Zhang L, Papasian CJ, Deng HW (2011). Comparative studies of *de novo* assembly tools for next-generation sequencing technologies. Bioinformatics, 27:15, 2031-2037.

123. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, McCalmon S, Hagerman RJ, Tassone F, Hagerman PJ (2013). Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Research, 23:1, 121-128.

124. Ma KY, Qiu GF, Feng JB, Li JL (2012). Transcriptome analysis of the oriental river prawn, *Macrobrachium nipponense* using 454 pyrosequencing for discovery of genes and markers. PLoS One, 7:6, e39727.

125. Ma KY, Lin JY, Guo SZ, Chen Y, Li JL, Qiu GF (2013). Molecular characterization and expression analysis of an insulin-like gene from the androgenic gland of the oriental river prawn, *Macrobrachium nipponense*. General and Comparative Endocrinology, 185:-, 90-96.

126. MacRae TH (2010). Gene expression, metabolic regulation and stress tolerance during diapause. Cellular and Molecular Life Sciences, 67:14, 2405-2424.

127. Maderspacher F (2008). Genomics: an inordinate fondness for beetles. Current Biology, 18:11, R466-468.

128. Magwene PM, Willis JH, Kelly JK (2011). The statistics of bulk segregant analysis using next generation sequencing. PLoS Computational Biology, 7:11, e1002255.

129. Mair G (2007). Genetics and breeding in seed supply for inland aquaculture. In. Assessment of Freshwater Fish Seed Resources for Sustainable Aquaculture. pp. 628.

130. Maneeruttanarungroj C, Pongsomboon S, Wuthisuthimethavee S, Klinbunga S, Wilson KJ, Swan J, Li Y, Whan V, Chu KH, Li CP, Tong J, Glenn K, Rothschild M, Jerry D, Tassanakajon A (2006). Development of polymorphic expressed sequence tag-derived microsatellites for the extension of the genetic linkage map of the black tiger shrimp (*Penaeus monodon*). Animal Genetics, 37:4, 363-368.

131. Maniatsi S, Baxevanis AD, Kappas I, Deligiannidis P, Triantafyllidis A, Papakostas S, Bougiouklis D, Abatzopoulos TJ (2011). Is polyploidy a persevering accident or an adaptive evolutionary pattern? The case of the brine shrimp *Artemia*. Molecular Phylogenetics and Evolution, 58:2, 353-364.

132. Manolakou P, Lavranos G, Angelopoulou R (2006). Molecular patterns of sex determination in the animal kingdom: a comparative study of the biology of reproduction. Reproductive Biology and Endocrinology, 4:59.

133. Mantovani B, Cesari M, Luchetti A, Scanabissi F (2008). Mitochondrial and nuclear DNA variability in the living fossil *Triops cancriformis* (Bosc, 1801) (Crustacea, Branchiopoda, Notostraca). Heredity, 100:5, 496-505.

134. Marcoval MA, Pan J, Tang YZ, Gobler CJ (2013). The ability of the branchiopod, *Artemia salina*, to graze upon harmful algal blooms caused by *Alexandrium fundyense*, *Aureococcus anophagefferens*, and *Cochlodinium polykrikoides*. Estuarine Coastal and Shelf Science, 131:235-244.

135. Mardis ER (2013). Next-generation sequencing platforms. Annual Review of Analytical Chemistry, 6:-, 287-303.

136. Marinho-Soriano EA, C.A.A.; Trigueiro, T.G.; Pereira, D.C.; Carneiro, M.A.A.; Camara, M.R. (2011). Bioremediation of aquaculture wastewater using macroalgae and *Artemia*. International Biodeterioration & Biodegradation, 65:1, 253-257.

137. Matson CK, Zarkower D (2012). Sex and the singular DM domain: insights into sexual regulation, evolution and plasticity. Nature Reviews Genetics, 13:3, 163-174.

138. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, Goodwin L, Woyke T, Lapidus A, Klenk HP, Cottingham RW, Kyrpides NC (2012). The fast changing landscape of sequencing technologies and their impact on microbial genome assemblies and annotation. PLoS One, 7:12, e48837.

139. Mendonça MAC, Carvalho CR, Clarindo WR (2010). DNA content differences between male and female chicken (*Gallus gallus domesticus*) nuclei and Z and W chromosomes resolved by image cytometry. Journal of Histochemistry and Cytochemistry, 58:3, 229-235.

140. Meyer JS, Suedkamp MJ, Morris JM, Farag AM (2006). Leachability of protein and metals incorporated into aquatic invertebrates: are species and metals-exposure history important? Archives of Environmental Contamination and Toxicology, 50:1, 79-87.

141. Michelmore RW, Paran I, Kesseli RV (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proceedings of the National Academy of Sciences of the United States of America, 88:21, 9828-9832.

142. Miller JR, Koren S, Sutton G (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95:6, 315-327.

143. Mitchell B, Crews ST (2002). Expression of the *Artemia trachealess* gene in the salt gland and epipod. Evolution & Development, 4:5, 344-353.

144. Muñoz J, Green AJ, Figuerola J, Amat F, Rico C (2009). Characterization of polymorphic microsatellite markers in the brine shrimp *Artemia* (Branchiopoda, Anostraca). Molecular Ecology Resources, 9:2, 547-550.

145. Muñoz J, Pacios F (2010). Global biodiversity and geographical distribution of diapausing aquatic invertebrates: the case of the cosmopolitan brine shrimp, *Artemia* (Branchiopoda, Anostraca). Crustaceana, 83:4, 465-480.

146. Nagaraju J, Gopinath G, Sharma V, Shukla JN (2014). Lepidopteran sex determination: a cascade of surprises. Sexual Development, 8:1-3, 104-112.

147. Nambu F, Tanaka S, Nambu Z (2007). Inbred strains of brine shrimp derived from *Artemia franciscana*: lineage, RAPD analysis, life span, reproductive traits and mode, adaptation, and tolerance to salinity changes. Zoological Science, 24:2, 159-171.

148. Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, Shan Z, Haaf T, Shimizu N, Shima A, Schmid M, Schartl M (2002). A duplicated copy of *DMRT1* in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*. Proceedings of the National Academy of Sciences of the United States of America, 99:18, 11778-11783.

149. *Lepeophtheirus salmonis* (salmon louse). NCBI.

http://www.ncbi.nlm.nih.gov/genome/2713

150. FTP directory /refseq/release at ftp.ncbi.nlm.nih.gov. NCBI

ftp://ftp.ncbi.nlm.nih.gov/refseq/release

151. Sequence assembly and alignment. (2009) Noonan J.

http://www.gersteinlab.org/courses/452/09-spring/pdf/SeqAssembly.pdf

152. New data release: *Arabidopsis* assembly offers glimpse of *de novo* SMRT sequencing for larger genomes. (2014) Pacific Biosciences of California, Inc.

http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html

153. Pannell JR (2008). Consequences of inbreeding depression due to sex-linked loci for the maintenance of males and outcrossing in branchiopod crustaceans. Genetics Research, 90:1, 73-84.

154. Pareek CS, Smoczynski R, Tretyn A (2011). Sequencing technologies and genome sequencing. Journal of Applied Genetics, 52:4, 413-435.

155. Parnes S, Khalaila I, Hulata G, Sagi A (2003). Sex determination in crayfish: are intersex *Cherax quadricarinatus* (Decapoda, Parastacidae) genetically females? Genetical Research, 82:2, 107-116.

156. Parra G, Bradnam K, Korf I (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23:9, 1061-1067.

157. Parraguez M, Gajardo G, Beardmore JA (2009). The New World *Artemia* species *A. franciscana* and *A. persimilis* are highly differentiated for chromosome size and heterochromatin content. Hereditas, 146:2, 93-103.

158. Perez F, Erazo C, Zhinaula M, Volckaert F, Calderon J (2004). A sex-specific linkage map of the white shrimp *Penaeus* (*Litopenaeus*) *vannamei* based on AFLP markers. Aquaculture, 242:1-4, 105-118.

159. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, Robasky K, Zaranek AW, Lee JH, Ball MP, Peterson JE, Perazich H, Yeung G, Liu J, Chen L, Kennemer MI, Pothuraju K, Konvicka K, Tsoupko-Sitnikov M, Pant KP, Ebert JC, Nilsen GB, Baccash J, Halpern AL, Church GM, Drmanac R (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature, 487:7406, 190-195.

160. Piferrer F (2013). Epigenetics of sex determination and gonadogenesis. Developmental Dynamics, 242:4, 360-370.

161. Pilla EJS, Beardmore JA (1994). Genetic and morphometric differentiation in Old World bisexual species of *Artemia* (the brine shrimp). Heredity, 73:47-56.

162. Pruitt KD, Tatusova T, Maglott DR (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35:Database issue, D61-65.

163. Qiao H, Fu H, Jin S, Wu Y, Jiang S, Gong Y, Xiong Y (2012). Constructing and random sequencing analysis of normalized cDNA library of testis tissue from oriental river prawn (*Macrobrachium nipponense*). Comparative Biochemistry and Physiology Part D, Genomics & Proteomics, 7:3, 268-276.

164. Qiu Z, Tsoi SC, MacRae TH (2007). Gene expression in diapause-destined embryos of the crustacean, *Artemia franciscana*. Mechanisms of Development, 124:11-12, 856-867.

165. Quarrie SAL-J, V.; Kovačević, D.; Steed, A.; Pekić, S., (1999). Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. Journal of Experimental Botany, 50:337, 1299-1306.

166. Rajasree SRR, Kumar VG, Abraham LS, Manoharan N (2011). Assessment on the toxicity of engineered nanoparticles on the lifestages of marine aquatic invertebrate *Artemia salina*. International Journal of Nanoscience, 10:4-5, 1153-1159.

167. Rees DJ, Dufresne F, Glémet H, Belzile C (2007). Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. Genome, 50:2, 151-158.

168. Rees DJ, Belzile C, Glemet H, Dufresne F (2008). Large genomes among caridean shrimp. Genome, 51:2, 159-163.

169. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature, 463:7284, 1079-1098.

170. Reuben M, Lin R (2002). Germline X chromosomes exhibit contrasting patterns of histone H3 methylation in Caenorhabditis elegans. Dev Biol, 245:1, 71-82.

171. Rheinsmith EL, Hinegard R, Bachmann K (1974). Nuclear DNA amounts in crustacea. Comparative Biochemistry and Physiology, Part B, 48:3B, 343-348.

172. Robbins HM, Van Stappen G, Sorgeloos P, Sung YY, MacRae TH, Bossier P (2010). Diapause termination and development of encysted *Artemia* embryos: roles for nitric oxide and hydrogen peroxide. Journal of Experimental Biology, 213:9, 1464-1470.

173. Rode NO, Charmantier A, Lenormand T (2011). Male-female coevolution in the wild: evidence from a time series in *Artemia franciscana*. Evolution, 65:10, 2881-2892.

174. Ross JS, Cronin M (2011). Whole cancer genome sequencing by next-generation methods. American Journal of Clinical Pathology, 136:4, 527-539.

175. Routtu J, Jansen B, Colson I, De Meester L, Ebert D (2010). The first-generation *Daphnia* magna linkage map. BMC Genomics, 11:508.

176. Ruan J, Jiang L, Chong Z, Gong Q, Li H, Li C, Tao Y, Zheng C, Zhai W, Turissini D, Cannon CH, Lu X, Wu C-I (2013). Pseudo-Sanger sequencing: massively parallel production of long and near error-free reads using NGS technology. BMC Genomics, 14:711.

177. Saccone G, Salvemini M, Polito LC (2011). The transformer gene of *Ceratitis capitata*: a paradigm for a conserved epigenetic master regulator of sex determination in insects. Genetica, 139:1, 99-111.

178. Salz HK (2011). Sex determination in insects: a binary decision based on alternative splicing. Current Opinion in Genetics & Development, 21:4, 395-400.

179. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marcais G, Pop M, Yorke JA (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Research, 22:3, 557-567.

180. Sander JD, Joung JK (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nature Biotechnology, 32:4, 347-355.

181. Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74:12, 5463-5467.

182. Schiex TM, A, Rouzé, P (2001). EuGene: an eucaryotic gene finder that combines several sources of evidence. Computational Biology. O. Gascuel and M-F. Sagot ed. Berlin: Springer. pp. 111-125.

183. Seidler RJ, Mandel M (1971). Quantitative aspects of deoxyribonucleic acid renaturation: base composition, state of chromosome replication, and polynucleotide homologies. Journal of Bacteriology, 106:2, 608-614.

184. Shen XY, Kwan HY, Thevasagayam NM, Prakki SR, Kuznetsova IS, Ngoh SY, Lim Z, Feng F, Chang A, Orban L (2014). The first transcriptome and genetic linkage map for Asian arowana. Molecular Ecology Resources, 14:3, 622-635.

185. Shin SC, Ahn do H, Kim SJ, Lee H, Oh TJ, Lee JE, Park H (2013). Advantages of single-molecule real-time sequencing in high-GC content genomes. PLoS One, 8:7, e68824.

186. Shuster SM, Levy L (1999). Sex-linked inheritance of a cuticular pigmentation marker in the marine isopod, *Paracerceis sculpta* Holmes (Crustacea : Isopoda : Sphaeromatidae). Journal of Heredity, 90:2, 304-307.

187. Siegismund HR (2002). Disparity in population differentiation of sex-linked and autosomal variation in sibling species of the *Jaera albifrons* (Isopoda) complex. Journal of Heredity, 93:6, 432-439.

188. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19:6, 1117-1123.

189. Smith CA, Roeszler KN, Ohnesorg T, Cummins DM, Farlie PG, Doran TJ, Sinclair AH (2009). The avian Z-linked gene *DMRT1* is required for male sex determination in the chicken. Nature, 461:7261, 267-271.

190. Soltanian S (2007). Protection of gnotobiotic *Artemia* against *Vibrio campbellii* using baker's yeast strains and extracts . PhD. Belgium: Ghent University. 201 p.

191. Soto-Jimenez MF, Arellano-Fiore C, Rocha-Velarde R, Jara-Marini ME, Ruelas-Inzunza J, Paez-Osuna F (2011). Trophic transfer of lead through a model marine four-level food chain: *Tetraselmis suecica*, *Artemia franciscana*, *Litopenaeus vannamei*, and *Haemulon scudderi*. Archives of Environmental Contamination and Toxicology, 61:2, 280-291.

192. Staelens J, Rombaut D, Vercauteren I, Argue B, Benzie J, Vuylsteke M (2008). High-density linkage maps and sex-linked markers for the black tiger shrimp (*Penaeus monodon*). Genetics, 179:2, 917-925.

193. Stefani R (1963). La digametia femminile in *Artemia salina* Leach e la constituzione del corredo cromosomico nei biotitic diploidi anfigonico e diploide partenogenético. Caryologia, 16, :-, 625-636.

194. Stemple DL (2013). So, you want to sequence a genome. Genome Biology, 14:7, 128.

195. Sterck L, Billiau K, Abeel T, Rouze P, van de Peer Y (2012). ORCAE: online resource for community annotation of eukaryotes. Nature Methods, 9:11, 1041-1041.

196. Stillman JH, Colbourne JK, Lee CE, Patel NH, Phillips MR, Towle DW, Eads BD, Gelembuik GW, Henry RP, Johnson EA, Pfrender ME, Terwilliger NB (2008). Recent advances in crustacean genomics. Integrative and Comparative Biology, 48:6, 852-868.

197. Sun Y. SW-Q, Zhong Y-C., Zhang R-S., Abatzopoulos T.J., Chen R-Y. (1999). Diversity and genetic differentiation in *Artemia* species and populations detected by AFLP markers. International Journal of Salt Lake Research, 8:4, 341-350.

198. Tarailo-Graovac M, Chen N (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics 4:10, 1-14.

199. Tatarazako N, Oda S, Watanabe H, Morita M, Iguchi T (2003). Juvenile hormone agonists affect the occurrence of male *Daphnia*. Chemosphere, 53:8, 827-833.

200. Tomaszkiewicz M, Smolarz K, Wolowicz M (2010). Heterogamety in the Baltic Glacial Relict *Saduria Entomon* (Isopoda: Valvifera). Journal of Crustacean Biology, 30:4, 757-761.

201. Torrentera L, Abreu-Grobois FA (2002). Cytogenetic variability and differentiation in *Artemia* (Branchiopoda: Anostraca) populations from the Yucatán Peninsula, Mexico. Hydrobiologia, 486:1, 303-314.

202. Toyota K, Kato Y, Sato M, Sugiura N, Miyagawa S, Miyakawa H, Watanabe H, Oda S, Ogino Y, Hiruta C, Mizutani T, Tatarazako N, Paland S, Jackson C, Colbourne JK, Iguchi T (2013). Molecular cloning of *doublesex* genes of four Cladocera (water flea) species. BMC Genomics, 14:1, 239.

203. Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25:9, 1105-1111.

204. Triantaphyllidis GV, Criel GRJ, Abatzopoulos TJ, Thomas KM, Peleman J, Beardmore JA, Sorgeloos P (1997). International Study on *Artemia*: 57. Morphological and molecular characters suggest conspecificity of all bisexual European and North African *Artemia* populations. Marine Biology, 129:3, 477-487.

205. Triantaphyllidis GV, Abatzopoulos TJ, Sorgeloos P (1998). Review of the biogeography of the genus *Artemia* (Crustacea, Anostraca). Journal of Biogeography, 25:2, 213-226.

206. RNAprep: Trizol combined with columns. (2008) Untergasser A

www.untergasser.de/lab

207. Vainola R (1998). A sex-linked locus (Mpi) in the opossum shrimp *Mysis relicta*: implications for early postglacial colonization history. Heredity, 81:-, 621-629.

208. Valverde JR, Batuecas B, Moratilla C, Marco R, Garesse R (1994). The complete mitochondrial DNA sequence of the crustacean *Artemia franciscana*. Journal of Molecular Evolution, 39:4, 400-408.

209. van Heesch S, Kloosterman WP, Lansu N, Ruzius F, Levandowsky E, Lee CC, Zhou S, Goldstein S, Schwartz DC, Harkins TT, Guryev V, Cuppen E (2013). Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC Genomics, 14:257 1-11.

210. Van Nieuwerburgh F, Thompson RC, Ledesma J, Deforce D, Gaasterland T, Ordoukhanian P, Head SR(2012). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Research, Location.

211. van Ooijen JW (2006). JoinMap® 4, Software for the calculation of genetic linkage maps in experimental populations. Wageningen, Kyazma B.V. 63 p.

212. Vaughn JC (1977). DNA reassociation kinetic analysis of brine shrimp, *Artemia salina*. Biochemical and Biophysical Research Communications, 79:2, 525-531.

213. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A,

Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X (2001). The sequence of the human genome. Science, 291:5507, 1304-1351.

214. Ventura T, Aflalo ED, Weil S, Kashkush K, Sagi A (2011). Isolation and characterization of a female-specific DNA marker in the giant freshwater prawn *Macrobrachium rosenbergii*. Heredity, 107:5, 456-461.

215. Ventura T, Sagi A (2012). The insulin-like androgenic gland hormone in crustaceans: From a single gene silencing to a wide array of sexual manipulation-based biotechnologies. Biotechnology Advances, 30:6, 1543-1550.

216. Ventura TA, E.D. Sagi, A. (2009). Future prospects of crustacean monosex culture: could giant prawn monosex culture benefit from the discovery of an insulin-like factor? Aquaculture Europe, 34:1, 30-31.

217. Vergilino R, Belzile C, Dufresne F (2009). Genome size evolution and polyploidy in the *Daphnia pulex* complex (Cladocera: Daphniidae). Biological Journal of the Linnean Society, 97:1, 68-79.

218. Verhulst EC, van de Zande L, Beukeboom LW (2010). Insect sex determination: it all evolves around *transformer*. Current Opinion in Genetics & Development, 20:4, 376-383.

219. Vezzi F, Narzisi G, Mishra B (2012). Feature-by-feature--evaluating *de novo* sequence assembly. PLoS One, 7:2, e31002.

220. von Reumont BM, Jenner RA, Wills MA, Dell'Ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A, Niehuis O, Meusemann K, Misof B (2012). Pancrustacean phylogeny in the light of new phylogenomic data: Support for Remipedia as the possible sister group of Hexapoda. Molecular Biology and Evolution, 29:3, 1031-1045.

221. Voorrips RE (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. Journal of Heredity, 93:1, 77-78.

222. Voronin DA, Kiseleva EV (2007). Functional role of proteins containing ankyrin repeats. Tsitologiia, 49:12, 989-999.

223. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Friters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995). AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, 23:21, 4407-4414.

224. Vuylsteke M, Peleman JD, van Eijk MJT (2007). AFLP-based transcript profiling (cDNA-AFLP) for genome-wide expression analysis. Nature Protocols, 2:6, 1399-1413.

225. Vuylsteke M, Peleman JD, van Eijk MJT (2007). AFLP technology for DNA fingerprinting. Nature Protocols, 2:6, 1387-1398.

226. Wang W, Tian Y, Kong J, Li X, Liu X, Yang C (2012). Integration genetic linkage map construction and several potential QTLs mapping of Chinese shrimp (*Fenneropenaeus chinensis*) based on three types of molecular markers. Russian Journal of Genetics, 48:4, 422-434.

227. Weeks SC, Benvenuto C, Sanderson TF, Duff RJ (2010). Sex chromosome evolution in the clam shrimp, *Eulimnadia texana*. Journal of Evolutionary Biology, 23:5, 1100-1106.

228. Wilson K, Li YT, Whan V, Lehnert S, Byrne K, Moore S, Pongsomboon S, Tassanakajon A, Rosenberg G, Ballment E, Fayazi Z, Swan J, Kenway M, Benzie J (2002). Genetic mapping of the black tiger shrimp *Penaeus monodon* with amplified fragment length polymorphism. Aquaculture, 204:3-4, 297-309.

229. Yandell M, Ence D (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13:5, 329-342.

230. You EM, Liu KF, Huang SW, Chen M, Groumellec ML, Fann SJ, Yu HT (2010). Construction of integrated genetic linkage maps of the tiger shrimp (*Penaeus monodon*) using microsatellite and AFLP markers. Animal Genetics, 41:4, 365-376.

231. Zdobnov EM, Apweiler R (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17:9, 847-848.

232. Zeng H, Song WQ, Chen RY (2004). Study of a DNA sequence from brine shrimp *Artemia* containing a novel DM domain. Shi Yan Sheng Wu Xue Bao, 37:5, 423-427.

233. Zhang HX, Luo QB, Sun J, Liu F, Wu G, Yu J, Wang WW (2013). Mitochondrial genome sequences of *Artemia tibetiana* and *Artemia urmiana*: assessing molecular changes for high plateau adaptation. Science China-Life Sciences, 56:5, 440-452.

234. Zhang LS, Yang CJ, Zhang Y, Li L, Zhang XM, Zhang QL, Xiang JH (2007). A genetic linkage map of Pacific white shrimp (*Litopenaeus vannamei*): sex-linked microsatellite markers and high recombination rates. Genetica, 131:1, 37-49.

235. Zhang Y, Zhang S, Liu Z, Zhang L, Zhang W (2013). Epigenetic modifications during sex change repress gonadotropin stimulation of *cyp19a1a* in a teleost ricefield eel (*Monopterus albus*). Endocrinology, 154:8, 2881-2890.

236. Zhang Z-Q (2011). Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness Zootaxa, -: 3148 1-237.

237. Zhong D, Pai A, Yan G (2004). AFLP-based genetic linkage map for the red flour beetle (*Tribolium castaneum*). The Journal of Heredity, 95:1, e68374.

238. Zhou R, Yang F, Chen DF, Sun YX, Yang JS, Yang WJ (2013). Acetylation of chromatin-associated Histone H3 Lysine 56 inhibits the development of encysted *Artemia* embryos. PLoS One, 8:6, e68374.

239. Zhou Y, Bizzaro JW, Marx KA (2004). Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. BMC Genomics, 5:95.

Curriculum Vitae

Personal information

First & last name: Stephanie De Vos
Gender: Female
Date & place of birth: 23/12/1983, Jette, Belgium
Nationality: Belgian

Contact details

Permanent address: Hundelgemsesteenweg 157, 9000 Gent
Email address: stephanie.devos@ugent.be
Cellular phone no.: 0032/497.66.53.55

Education

PhD (2009 – 2014):

Ghent University, Faculty of Bioscience Engineering, Laboratory of Aquaculture & Artemia Reference Center (Prof. Dr. ir. Peter Bossier)

Flemish Institute of Biotechnology, VIB-PSB, quantitative genomics (Dr. ir. Marnik Vuylsteke)

Dissertation title: “Genomic tools and sex determination in the extremophile brine shrimp *Artemia franciscana*”.

Successful completion of the Doctoral Training Program (soft skills, seminars and specialist courses) organized by the Doctoral School of Bioscience Engineering.

Master in Bioscience Engineering, option Agronomic Engineering (2001 – 2008):

Ghent University, Faculty of Bioscience Engineering

Dissertation title: “Characterization, pathogenicity and integrated control of Cuban *Rhizoctonia solani*-isolates on bean”. The dissertation was completed at the group of Crop protection (Prof. Dr. ir. Monica Höfte), Ghent University and included one month of laboratory work at the Instituto Biotechnologico de Las Plantas (IBP), Universidad Central de Las Villas, Cuba.

Obtained with distinction.

Training sessions attended

- Applied bioinformatics in Plant Sciences (Athens, Greece, 2011)
- COST Training School on Next Generation Sequencing (Uppsala, Sweden, 2011)
- MG4U Summer Course: Marine Evolutionary & Ecological Genomics (Roscoff, France, 2011)
- Introductory session on programming with Bioperl (VIB Bits, Zwijnaarde, Belgium, 2011)
- Perl and Bioperl Introductionary Course (VIB Bits, Zwijnaarde, Belgium, 2011)
- UCSC Genome Browser Training (VIB Bits, Zwijnaarde, Belgium, 2011)
- Advanced Academic English: Writing Skills for (Bioscience) Engineering (Ghent University, Belgium, 2012)

Scientific activities

- Posters: 11th VLIZ Young Marine Scientists' Day (Bruges, Belgium, 2011): “A first AFLP-based linkage map and sex-linked markers for *Artemia”*

Genetics, epigenetics and evolution of sex chromosomes (Institut Jaques Monod, Paris, France, 2011): “A first AFLP-based linkage map and sex-linked markers for *Artemia”*

- Talks: PhD symposium (Gent, Belgium, 2012): “A first AFLP-based linkage map and *de novo* assembled genome sequence for *Artemia”.*

VIB seminar, session “NGS: the state of affairs in VIB” (Blankenberge, Belgium, 2012): “*Artemia de novo* genome sequence and sex-linked SNP`s by bulked segregant analysis”.

Larvi 2013, 6th fish & shellfish larviculture symposium, International workshop, session ”Brine shrimp *Artemia* as a model organism in life sciences research” (Gent, Belgium, 2013): “*Artemia* Genomics”.

-Teaching: Supervision Master thesis Adeniyi Racheal Tolani, Ugent (2011).

Practical exercises `PCR-RFLP of Artemia cysts`, Master in Aquaculture, Ugent (2011)

Genetics course, Master in Aquaculture, Ugent (2012 & 2013).

- Congresses attended without presenting a talk or poster:

4th From PhD to Job Market conference (VUB, Brussels, Belgium, 2011).

Language skills

- Mother tongue: Dutch
- Very fluent: English, French
- Basic knowledge: Spanish, German

Personal interests

Most of all, I am interested in the areas of Genomics, Aquaculture and (sub)tropical agriculture.

Publications

De Vos S, Bossier P, Van Stappen G, Vercauteren I, Sorgeloos P, Vuylsteke M (2013) A first AFLP-based genetic linkage map for brine shrimp *Artemia franciscana* and its application in mapping the sex locus. PLoS One, 8 (3), e57585.

Nerey Y, Pannecoucque J, Hernandez HP , Diaz M, Espinosa R, **De Vos, S**, Van Beneden S, Herrera L, Höfte M (2010) *Rhizoctonia* spp. Causing Root and Hypocotyl Rot in Phaseolus vulgaris in Cuba. Journal of Phytopathology, 158 (4), 236-243.