Dissertation

submitted to the

Combined Faculties of Natural Sciences and Mathematics

of the

Ruperto-Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

M. Sc. Julian Gutekunst
born in Herrenberg, Germany

Oral examination: _____

Clonal genome evolution of the marbled crayfish,

Procambarus virginalis

Referees

Prof. Dr. Frank Lyko

Prof. Dr. Benedikt Brors

Abstract

Marbled crayfish (*Procambarus virginalis*) are the only freshwater crayfish known to reproduce by cloning (apomictic parthenogenesis). Notably, among genetically identical offspring raised in the same environment, distinct phenotypic differences can be observed. These unique characteristics establish the marbled crayfish as a particularly interesting laboratory model. Additionally, parthenogenetic reproduction enables the marbled crayfish to rapidly spread and form stable populations, which poses a serious threat in many freshwater habitats. A further understanding of this organism requires the accessibility of its 3.5 Gbp large genome sequence.

This doctoral thesis provides the first *de novo* genome assembly of the marbled crayfish. Multiple shotgun and long jumping distance libraries were generated from one individual female, with a single base coverage of over 100x. Sequencing data was used for a first genome assembly with a length weighted median scaffold size (N50) of over 40 kbp. The estimated genome wide heterozygosity rate of 0.53% is substantially higher compared to other arthropod genomes. Transcriptome data enabled the refinement of genetic structures. Eventually, a total of 87.8% complete and 7.4% fragmented single-copy arthropod orthologs were identified using the benchmarking software BUSCO. Single nucleotide variations were analyzed to verify clonality in geographically isolated populations. Results indicate an evolution from a single origin. Moreover, detailed insights into genotype distributions support the theory of asexual speciation by autopolyploidization. Comparison of three *Procambarus* species indicates detectable genetic separation between marbled crayfish and the closest relative *Procambarus fallax*. Automatic annotation of 21,000 genes using the annotation pipeline MAKER provides a detailed overview of genetic features. For example, a cellulase gene was identified which potentially plays a key role in omnivorousness. Genomic data and several online services are provided by a central web resource (http://marmorkrebs.dkfz.de).

This thesis provides detailed genetic insights into the unknown but very versatile order of decapod crustaceans. Considered economically and ecologically relevant keystone species, a representative genome sequence provides an important resource for future research.

Kurzzusammenfassung

Marmorkrebse (*Procambarus virginalis*) sind die bislang einzig bekannten Süßwasserkrebse, die sich ausschließlich per Jungfernzeugung (apomiktische Parthenogenese) fortpflanzen. Interessanterweise kann man unter genetisch identischen Tieren, die unter denselben Bedingungen aufgewachsen sind, phänotypische Veränderungen beobachten. Diese einzigartigen Charakteristiken etablieren den Marmorkrebs als einen besonders interessanten Modellorganismus für biologische Studien. Zudem erlaubt der Reproduktionsmodus eine rasche Verbreitung und Gründung von beständigen Populationen. Das macht sie zu einer besonders großen Gefahr in vielen Süßwassersystemen. Um den Marmorkrebs besser zu verstehen bedarf es daher Kenntnis seiner 3,5 Gigabasen großen Genomsequenz.

Im Rahmen dieser Doktorarbeit wurde eine erste Genomsequenz des Marmorkrebses etabliert. Dafür wurden von einem weiblichen Tier verschiedene Sequenzbibliotheken, mit einer über 100-fachen genomweiten Nukleotidabdeckung, generiert. Die Sequenzierungsdaten ermöglichten die Assemblierung einer Genomsequenz mit einem gewichteten Median der Sequenzlängen (N50) von über 40 Kilobasen. Die genomweite Heterozygosität im Marmorkrebs wurde auf 0.53% geschätzt. Damit ist sie deutlich höher im Vergleich zu anderen Arthropoden. Zusätzliche Transkriptominformationen erlaubten die Korrektur von genetischen Strukturen. Die Qualität der resultierenden Assemblierung wurde mittels einer universellen orthologen Gendatenbank bewertet (BUSCO). So konnten 87.8% komplette und 7.4% fragmentierte Gene gefunden werden. Die Analyse von Einzelnukleotidvariationen zeigte eine klonale Evolution in geografisch isolierten Marmorkrebs Populationen. Die Ergebnisse deuten auf eine Abstammung von einem einzigen Ursprungstier. Zudem wurden Einblicke in genomische Charakteristiken gewährt, wie etwa Triploidität und verschieden vorkommende Genotypen. Dabei war die Verteilung der Genotypen ein deutliches Indiz für eine Speziesbildung durch Autopolyploidisierung. Dies wurde ebenso durch vergleichende Analysen verdeutlicht, die eine klar erkennbare genetische Separation des Marmorkrebses zu seinem nächsten Verwandten, *Procambarus fallax*, zeigten. Die automatische Genomannotation von über 21,000 Genen gab Einblicke in genomische Regionen. So konnte eine spezifische Cellulase identifiziert werden, welche möglicherweise eine wichtige Rolle in der Omnivorosität in Süßwasserkrebsen spielt. Genomische Daten und diverse Online Tools werden über ein zentrales Web System angeboten (http://marmorkrebs.dkfz.de).

Diese Doktorarbeit gewährt detaillierte genomische Einblicke in die bisher unbekannte aber sehr vielseitige Ordnung der Zehnfußkrebse (Dekapoden). Da diese als ökonomische und ökologische Schlüsselarten gelten bietet die repräsentative Genomsequenz ein wichtiges Werkzeug für zukünftige biologische Forschungen.

Contents

List of Figures

List of Tables

List of Abbreviations

5mC	5-methylcytosine
A	Adenine
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BUSCO	Benchmarking Universal Single-Copy Orthologs
C	Cytosine
CDS	(Protein) Coding sequence
cm	Centimetre
CpG	Cytosine Guanosine dinucleotide
CPU	Central processing unit
DBG	De Brujin graph
ddNTP	Dideoxynucleotide
DNA	Deoxyribonucleic acid
G	Guanine
Gbp	Gigabase pair ($\times 10^9$ bp)
GH9	Glycoside Hydrolase Family 9
GO	Gene ontology
HCS	HiSeq control software
HPC	High performance computing cluster
kbp	Kilobase pair ($\times 10^3$ bp)
LINE	Long interspersed nuclear element
LJD	Long jumping distance
MA	Marbled crayfish sample from Madagascar
Mbp	Megabase pair ($\times 10^6$ bp)
µg	Microgram
N	Ambiguous gap nucleotide
nt	Nucleotide
OLC	Overlap-Layout-Consensus
RTA	Real time analysis (software)
SBS	Sequencing by synthesis
SG	Shotgun (reads)
SINE	Short interspersed nuclear element
SMRT	Single molecule real time (sequencing)
SNP	Single nucleotide polymorphism
T	Thymine
TCGA	The Cancer Genome Atlas
TE	Transposable element
WGBS	Whole genome bisulfite sequencing
WGS	Whole genome sequencing

1 Introduction

With more than 640 different species, crayfish are considered to have an ecological impact in freshwater ecosystems (Crandall and Buhay, 2008). Particularly, they play an important part in the maintenance of aquatic habitats. Their large abundance and omnivorousness are required to balance aquatic populations. Furthermore, crayfish have commercial value in many countries. Not only freshwater crayfish, but many decapod crustaceans such as crabs, prawns, shrimps, and lobsters are an important nutritional source. Consequently, freshwater crayfish are considered an economical and ecological keystone species. It is rather surprising that genome sequences for these animals are still rare. Currently, the water flea (*Daphnia pulex*) and the sand flea (*Parhyale hawaiensis*) provide the only two published crustacean genomes (Colbourne et al., 2011; Kao et al., 2016). The marbled crayfish, which is named for the marbled texture on its carapace (Figure 1.1A), is a particularly interesting freshwater crayfish due to its relevant role as a potential model organism and its emergence as invasive species (Vogt, 2008b; Jones et al., 2009).

1.1 Marbled crayfish

First documented records of the marbled crayfish were in a German aquarium trade in the mid-1990s (Lukhaup, 2001; Lukhaup and Pekny, 2003; Vogt et al., 2004). There, aquarists discovered they could establish populations from only one individual female. Scientific observations on genetic markers confirmed clonal reproduction, resulting in genetically identical offspring (Figure 1.1A) (Scholtz et al., 2003; Vogt et al., 2015). More specifically, the marbled crayfish exhibits an all-female population which reproduces by obligatory apomictic parthenogenesis (Scholtz et al., 2003; Martin et al., 2007; Vogt et al., 2008). Genomic

characterization further revealed a triploid genotype with a (haploid) set of 92 chromosomes (Vogt et al., 2015; Martin et al., 2016).

Phylogenetic analyses placed the marbled crayfish in the genus *Procambarus* (Scholtz et al., 2003). Specifically, identical morphology and microsatellite markers revealed a close relationship to sexually reproducing *Procambarus fallax*, which is native to Florida and southern Georgia (Scholtz et al., 2003; Martin et al., 2010a). However, the exact phylogenetic placement remained unclear. Ultimately, it was assumed that the marbled crayfish is the parthenogenetic form of *P. fallax* (Martin et al., 2010a). Thus, the scientific name *Procambarus fallax* forma *virginalis* was used until further studies were conducted. However, several observations suggest the marbled crayfish as an independent asexual species (Vogt et al., 2015). For example, marbled crayfish were not found near the natural habitat of *P. fallax*. Moreover, isolated reproduction and substantial genetic and epigenetic differences between both species were observed (Vogt et al., 2015; Falckenhayn, 2017). Conclusively, it was proposed that the marbled crayfish evolved from *P. fallax* by asexual speciation (Vogt et al., 2015) and the name *Procambarus virginalis* was adapted, as it was already suggested by Martin et al. (2010a).

So far, only speculations about the evolution of parthenogenesis in marbled crayfish exist. Different forms and origins could already be dismissed (Martin et al., 2007; Vogt et al., 2004). It was shown that the marbled crayfish evolved from the morphological identical *Procambarus fallax* (Scholtz et al., 2003; Martin et al., 2010a; Vogt et al., 2015). In combination with further studies on genetic markers, an evolution by autopolyploidization was suggested (Vogt et al., 2015; Martin et al., 2016). Particularly, a macromutation event caused by a heat or cold shock in captive *P. fallax* was speculated (Vogt et al., 2015). It is rather likely that the marbled crayfish directly evolved from *Procambarus fallax*, without any evidence for hybridization.

Despite being morphologically identical to *P. fallax*, the marbled crayfish grows larger in body size and possesses higher fecundity (Vogt et al., 2015). Additionally, anthropogenic releases and stable populations in different habitats suggests high capabilities for environmental adaptation. First documented records of *P. virginalis* were in Germany (Marten et al., 2004; Chucholl et al., 2010; Martin et al., 2010b). Since then, populations were found in other European countries, such as Netherlands (Holdich and Pöckl, 2007), Italy (Marzano

et al., 2009), Sweden (Bohman et al., 2013), Hungary (Lőkkös et al., 2016), the Czech Republic (Patoka et al., 2016), Slovakia (Lipták et al., 2016), and Ukraine (Novitsky and Son, 2016) (Figure 1.1B). Interestingly, geographically distant releases in Madagascar (Jones et al., 2009) and Japan (Kawai and Takahata, 2010) were reported. The variety of regions show that marbled crayfish can tolerate a wide range of temperatures. In fact, studies on temperature conditions render it a potent invader of many freshwater habitats (Feria et al., 2011; Faulkes et al., 2012). Natural habitats include burrows, streams, lakes, rivers, and ponds. However, as it also resides in rice paddies, marbled crayfish pose an economical and ecological threat in Madagascar and Japan (Jones et al., 2009; Kawai and Takahata, 2010; Faulkes et al., 2012). There, the marbled crayfish nourishes from rice plants eventually destroying entire fields.

Figure 1.1: Morphology and worldwide distribution of the marbled crayfish. (A) Genetically identical marbled crayfish batch mates. Differences in size can be observed within the same clutch kept under identical laboratory conditions (scale bar: 1 cm). (B) Worldwide distribution of the marbled crayfish colored in red. Populations were reported in Germany (Marten et al., 2004; Chucholl et al., 2010; Martin et al., 2010b), Netherlands (Holdich and Pöckl, 2007), Italy (Marzano et al., 2009), Sweden (Bohman et al., 2013), Hungary (Lőkkös et al., 2016), the Czech Republic (Patoka et al., 2016), Slovakia (Lipták et al., 2016), Ukraine (Novitsky and Son, 2016), Madagascar (Jones et al., 2009), and Japan (Kawai and Takahata, 2010).

1.1.1 Marbled crayfish as a promising new research model

Model organisms play an essential role to elucidate general principles in biological research (Hedges, 2002). They often provide unique characteristics required for analyzing specific processes. In molecular biology, species such as *C. elegans*, *D. melanogaster*, or *A. thaliana* were well studied for the effects of genetic variation. These model organisms possess favorable traits such as large offspring sizes, high fertility, and easy handling and culturing. Most importantly, distinct phenotypic changes in different genotypes could be observed. Access to whole genome information facilitated genome wide studies. Nowadays, a vast amount of analyses is based on information provided by the genome sequence. Applications include studies on gene regulation and variant analyses. Nonetheless, most model organisms were designed for genetic analyses and are not convenient for other fields of research. This, as well as other factors, demand for new laboratory models.

Due to its unique genomic characteristics, the marbled crayfish was proposed as an interesting new model for various fields of research (Vogt, 2008b, 2011a). Furthermore, it shows favorable properties such as a large body size, high fecundity, and they are easy to handle and culture. Marbled crayfish were already used for studies in developmental biology, evolutionary biology, toxicology, and neuroanatomy (Rieger and Harzsch, 2008; Alwes and Scholtz, 2006; Vogt, 2007; Vilpoux et al., 2006). Furthermore, future perspectives for cancer treatment, aging, and stem cell research were proposed (Vogt, 2008a, 2009, 2011a, 2012). Remarkably, despite genetically identical offspring, phenotypic differences can be observed (Figure 1.1A). Especially in epigenetics, research on phenotypic plasticity without the influence of genetic variation is required. Hence, the marbled crayfish provides an interesting model to study epigenetic regulation 1.1A) (Vogt, 2008b; Vogt et al., 2015; Falckenhayn, 2017). Nevertheless, research on *P. virginalis* as a model organism is strictly limited without access to its complete genome sequence. Analysis on genetic markers, mtDNA, or microsatellites were already performed in previous studies. However, these analyses require extensive laboratory work. Whole genome information provides an essential tool to fully establish the marbled crayfish as a new laboratory model.

Moreover, the genome sequence of marbled crayfish contributes genomic insights in the subphylum of crustaceans. With more than 60,000 species, crustaceans are a major part in the phylum of arthropods. It is rather surprising that only two crustaceans, *Daphnia pulex*

and *Parhyle hawaiensis*, were sequenced so far (Colbourne et al., 2011; Kao et al., 2016). Additional genome information is crucial to further elucidate this genomically very unknown but important lineage.

1.2 The genome encodes heritable information

In eukaryotes, the genome comprises all inheritable traits of an organism. It provides essential information required to determine morphological and physiological traits. The entire genomic information, from protein coding regions to functional non-coding regions, is included in the genome. Therefore, it is built from deoxyribonucleic acid (DNA) and often structured in chromosomes. The number of chromosomes and copies (ploidy) can vary from organism to organism. DNA mainly consists of the four nucleotides Adenine (A), Cytosine (C), Thymine (T), and Guanine (G) which are integrated into a sugar-phosphate backbone. Generally, DNA strands are complemented by their reverse strands at the 3' end. Therefore, in Watson-Crick base pairing, A is pairs with T and C is pairs with G (and vice versa). The resulting structure resembles the characteristic double helix.

1.2.1 Sequencing DNA by Illumina short read sequencing

The process to determine nucleotide content in a DNA fragment is also known as DNA sequencing. One of the earliest successful (first generation) sequencing strategy was based on a chain-termination approach proposed by F. Sanger and colleagues (Sanger et al., 1977). Sanger sequencing starts by hybridizing single strands of target DNA fragments to specific primers (template). Next, standard nucleotides and fluorescent-labeled dideoxynucleotides (ddNTPs) are introduced and synthesized onto the template. Integration of ddNTPs ceases elongation such that all possible lengths of fragments are produced. This step is also known as chain termination. Then, capillary gel electrophoresis is used to separate fragments by size. Laser excitation releases fluorescence labels and, with the help of computational processing, the sequence can be inferred. Sequencing became popular with first commercial applications of Sanger sequencing. Since then, technologies were constantly improved. Specifically, the second generation was coined by Illumina. Until now, Illumina sequencing was used in a huge number of projects. Many *de novo* whole

genome sequencing projects rely on accurate high coverage short read sequencing. Nevertheless, further applications include variant analysis which depends on high coverage sequencing for an accurate identification of polymorphic sites (DePristo et al., 2011; Garrison and Marth, 2012). Moreover, modifications in library preparation allow for the detection of DNA methylation sites on a single base resolution (whole genome bisulfite sequencing) (Cokus et al., 2008; Falckenhayn, 2017). Although the experimental design may vary, the sequencing approach remains similar for all applications (Figure 1.2).

First, target DNA is isolated and sheared into short random fragments. As this resembles a shotgun firing pattern, this method is also known as shotgun sequencing. Next, specific adapters are ligated, allowing fragments to hybridize with flow cell fixed adapters. Synthesis of complement strands is initiated once fragments hybridize onto flow cells. Denaturation and washing away newly synthesized strands allows to keep only fixed DNA fragments on the flow cell. Next, clusters of double stranded sequences are generated by bridge amplification. Reverse strands are cleaved such that they can be washed away. The sequence is determined by synthesizing fluorescent-labeled nucleotides to each strand in the so-called sequencing by synthesis (SBS) step. Each cluster emits a colored fluorescence signal upon incorporation of a nucleotide. Signals are detected and measured by specific cameras and the sequence is automatically inferred. The same procedure can be performed for the reverse strand, resulting in paired-end reads. Eventually, a sequence library is created which required further computational (*in silico*) processing.

Latest proceedings in sequencing technologies focus on generating long reads of up to several Kilobases. This allows a more complete insight into the genomic structures of an organism. Moreover, certain limitations of current assembly strategies can be avoided. Especially, projects on large genomes immensely benefit from long read sequencing technologies. Details on these technologies as well as potential applications were further discussed in the outlook of this thesis (Section 4.2).

1.2.2 Bioinformatic approaches in genome assembly

Sequencing output for whole genome sequencing (obtained from Illumina) is limited to short reads of up to several hundred base pairs. Thus, genome assembly requires further computational processing. The assembly of short reads into long contiguous sequences depends

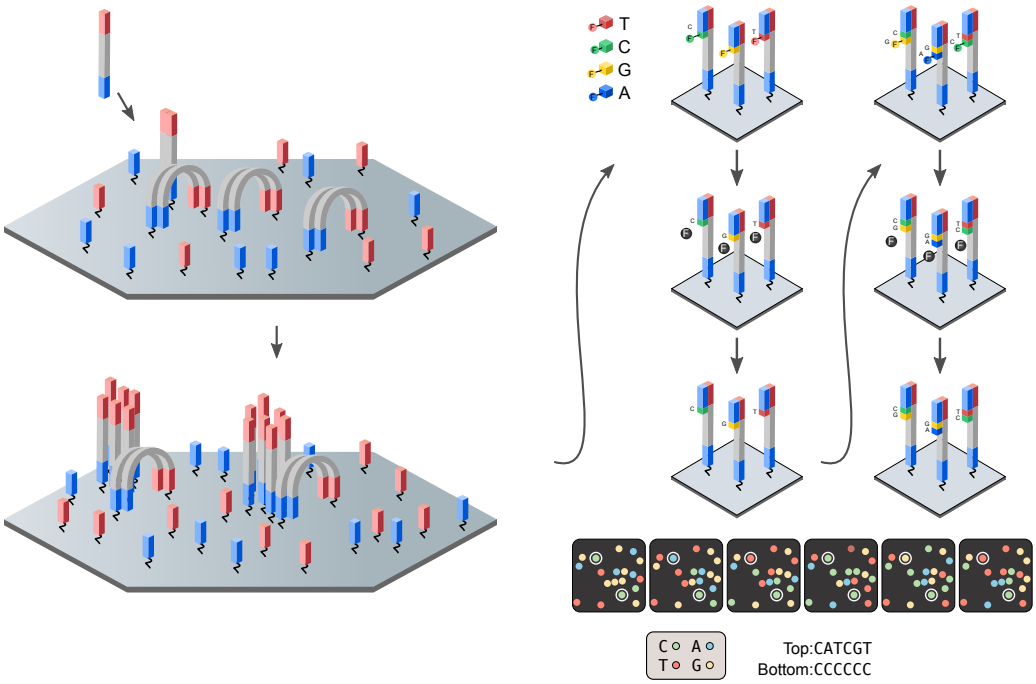

Figure 1.2: Schematic overview of Illumina short read sequencing. Illumina short read sequencing starts with ligation of DNA fragments onto flow cell fixed adapters. Bridge amplification of a single DNA fragment results in a dense cluster of sequences. Next, the sequencing by synthesis step is initiated. Therefore, fluorescence labeled nucleotides are synthesized onto single stranded fragments. Upon integration, clusters emit a distinct light for each nucleotide. Fluorescence emissions are measured by specific lasers and the sequence can automatically be inferred. The fluorescence labels are cleaved and washed away such that the process can be repeated until the complete sequence of a fragment is known. The image is adapted from Metzker (2010).

on an overlapping read design and high single nucleotide coverage. Computational approaches are used to create long sequences by connecting overlapping reads. Therefore, several methods have been developed. So far, graph based approaches proved to be the most successful strategies in terms of accuracy and efficiency (Miller et al., 2010). Graph based assemblers can be categorized by two mathematical concepts, the Overlap-Layout-Consensus method (OLC) and de Brujin graphs (DBG) (Miller et al., 2010).

Briefly, OLC assembly starts with a pairwise comparison and alignment of overlapping reads. This results in a basic graph layout, which further is resolved into consensus sequences (Myers, 1995). The OLC approach was primarily designed for the assembly of

Sanger sequencing reads. The most prominent implementation is provided by the Celera assembler which was used for the human genome project (Myers et al., 2000; Venter et al., 2001).

More frequent implementations rely on de Brujin graphs and path traversal algorithms (Luo et al., 2012; Zerbino and Birney, 2008; Simpson et al., 2009; Gnerre et al., 2011). De Brujin graph assemblies start with producing overlapping sub sequences of a defined length (k) from the initial sequencing reads, so called k-mers. K-mers are important for building and connecting the graph structure as sequence generation relies on the concept of k-mers overlapping by $k-1$ nucleotides. Each unique k-mer is represented as one vertex in the graph. Vertices are connected by edges, when k-mers are overlapping (Figure 1.3). Often, refinement is required to reduce complexity for path traversal. For example, bubbles emerge when alternative paths are present, as in the case of heterozygous alleles. Repetitive sequences create alternative paths in such way that a loop occurs (Figure 1.3). These structures need to be resolved as they will negatively affect the output. After constructing the graph, path traversal (eulerian or hamiltonian) is used to walk along edges and thus generating a continuous sequence (Figure 1.3). A sequence cannot further be extended when a vertex does not have any outgoing edges. In optimal conditions, each sequence represents one chromosome. However, the output generally remains fragmented due to unresolved difficulties in the graph structure such as ambiguous paths or sudden dead ends.

Over the last years assembly strategies were improved and modified to fit specific needs and allow the assembly of more complex genome structures. For example, adaptations exist to combine assembly strategies from various technologies or to resolve graph structures such that different haplotypes can be inferred (Zimin et al., 2013; Aguiar and Istrail, 2013; Kajitani et al., 2014).

1.2.3 Annotation – loci and features within a genome

De novo genome assembly results in a set of sequences without detailed information about loci of genetic features. Therefore, a variety of methods and tools have been developed to automatically infer gene locations, repeat structures, or other important features (Yandell and Ence, 2012). Especially, novel genomes depend on the knowledge of genetic loci, i.e.

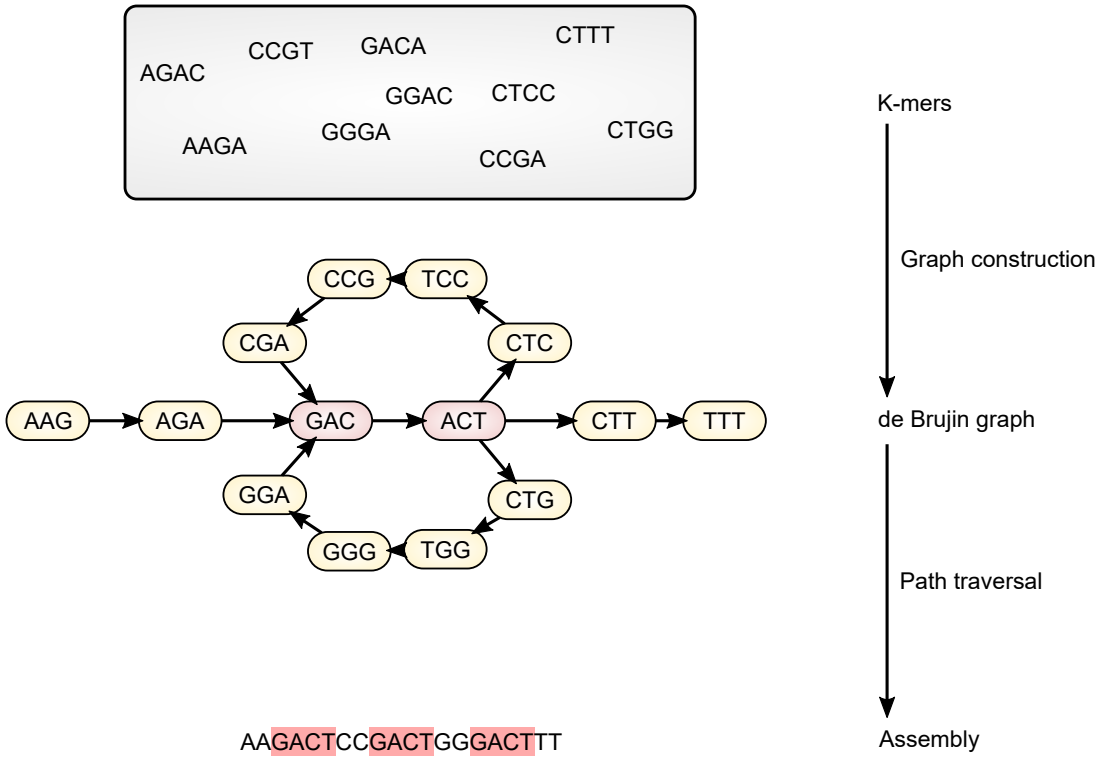

Figure 1.3: Schematic overview of de Brujin graphs in genome assembly. De Brujin graphs are commonly implemented in short read assemblers. Overlapping k-mers are built from sequencing reads with a fixed k-mer length. K-mers are then used as vertices in a graph. Edges represent connections of overlapping k-mers. The sequence is inferred by following edges in the graph and combining information from each vertex. Repetitive sequences (highlighted red in the last step) result from loops in the graph. The image is adapted from Berger et al. (2013).

to define gene characteristics (Eads et al., 2012), perform analyses on genomic features (Wang et al., 2014; Falckenhayn, 2017), or comparing gene conservation and synteny to other species (Braasch et al., 2016; Simakov et al., 2013; Machida et al., 2005).

Most important information is gained by gene locations. In eukaryotes, genes can be predicted by statistical models or by evidence from homologous genes, preferably from closely related organisms. Often, both methods are combined and a final gene prediction is based on a consensus model. Additional information provides detailed predictions about intron/exon structure or untranslated regions. Moreover, genetic features such as repeat structures or non-coding RNAs are predicted in a similar approach.

Nowadays, *de novo* genome annotation is fully automated. Pipelines exist for which often only homology databases are required. Certainly, further information enables more detailed predictions of other genetic features. Nevertheless, automated genome annotation often lacks in quality (Reese and Guigó, 2006). Therefore, analyses on specific genomic regions require careful practice. Annotations of larger genomes are published for review and quality control, also called manual curation. This allows research committees to correct specific regions by laboratory or literature research.

1.2.4 Genome projects

Eukaryotic genome sequencing projects started with the first complete genome of the baker's yeast *Saccharomyces cerevisiae* in 1996 (Goffeau et al., 1996). Within four years genetic model organisms followed, namely *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), and *Drosophila melanogaster* (Adams et al., 2000). Ultimately, a major goal in genome projects was to sequence the first human genome. Advancements in sequencing technologies and bioinformatical approaches have been made such that by 2001 a first draft of the human genome was published (McPherson et al., 2001; Venter et al., 2001). To date, it was the biggest and most expensive genome project. Continuous optimization and development of new strategies allowed for more cost-efficient projects. Soon, genome projects for various individuals from all major kingdoms were initiated. These often aimed to gain insights into genome evolution (Simakov et al., 2013; Albertin et al., 2015; Braasch et al., 2016), unique characteristics of an organism (Honeybee Genome Sequencing Consortium, 2006; Sanggaard et al., 2014; Wang et al., 2014), or about human diseases caused by an animal (Gulia-Nuss et al., 2016; Rosenfeld et al., 2016). Coordination and execution is either done by individual research groups or is part of a large genome sequencing initiative such as the i5K for arthropods (i5K Consortium, 2013), B10K for birds (Zhang et al., 2015), or G10K for vertebrates (Genome 10K Community of Scientists, 2009). Large initiatives often aim to obtain whole genome information of important representatives for a specific taxonomy.

Generally, genome projects always proceed in a similar practice regardless of the analyzed organism. A project is initiated with DNA extraction from the target animal. Suitable sequencing and assembly strategies are chosen, depending on prior knowledge about the

genomic characteristics. Sequencing is often based on short read sequencers, but recent advancements enable new sequencing technologies to emerge (discussed in Section 4.2). Assembly algorithms are further used and optimized depending on genome structure and complexity. Often, assembly optimizations are favorable depending on the research focus. Finally, genomic sequences are annotated using predictions based on homology and laboratory information.

1.3 Aims of this PhD thesis

Model organisms play an important role in molecular biology. Especially, whole genome information for organisms such as *C. elegans*, *D. melanogaster*, or *A. thaliana* facilitated genome wide analysis. Still, many laboratory models are heavily focused on genetics and thus are not suitable for many other applications in biological research. The marbled crayfish, *Probambarus virginalis*, possesses unique genomic characteristics. Obligate parthenogenesis and frequent releases allow rapid expansions in freshwater habitats. Consequently, the marbled crayfish is threatening local biodiversity and has negative ecological and economical impact. Nevertheless, it was proposed as a promising new model in various fields of biology. Its prevalent importance demands for genomic information.

This doctoral thesis focuses on establishing the genome sequence for the marbled crayfish. This includes sequencing on a variety of platforms, *de novo* assembly, and automatic annotation of genomic features. The genome assembly is ready to be used and distributed by a central web server. Until now, obligate parthenogenesis and triploidy were only analyzed by extensive laboratory work. Additional focus of this project includes a genome wide verification of genomic characteristics. Furthermore, globally distributed *P. virginalis* populations were analyzed for evolution of the parthenogenetic reproduction mode, ultimately confirming a single origin.

2 Materials and Methods

Genome projects require a variety of tools and hardware for data generation and analysis. This chapter comprises descriptions of the technical infrastructure and approaches for sequencing, assembly, and annotation of the marbled crayfish genome. Additionally, an overview of methods for variant detection and comparative analyses is given.

2.1 Technical infrastructure

Data generation and analysis was performed using different hardware settings and a variety of software. The following section provides an overview of hardware and software applications. Detailed parameter settings are explained in the respective paragraphs.

2.1.1 Hardware

Various applications require hardware settings dependent on resource demands and input size. Here, mainly three different hardware platforms were used. Small scripts and quick analyses were performed on a personal desktop computer equipped with four cores (one thread per core) and up to 32 GB of RAM. Computationally more demanding software was run on a high-performance computing cluster (HPC) with configurations of up to 10 nodes with 20 threads and 200 GB RAM. The web server is hosted on a powerful machine with four cores (two threads per core) and 24 GB RAM, which supplies sufficient resources for web based applications such as the Apollo genome annotation browser.

2.1.2 Software

Major software packages used in this project are listed below (Table 2.1). Each software is provided with a version number and reference. Software with multiple version numbers were either used in different stages of the project or were executed on different platforms. Minor packages and methods are provided in the respective paragraphs.

Table 2.1: Main software packages. Each software used in the scope of this project is provided with a version number and reference. Programs with multiple version numbers were executed in different computational environments. Minor software packages are not listed but described in the respective paragraphs.

Software	Version	Source
BBTools	34.72	Bushnell (2015)
BLAST	2.2.29+	Altschul et al. (1990)
Bowtie 2	2.2.6	Langmead and Salzberg (2012)
BUSCO	1.22 and 2.0	Simão et al. (2015)
ClustalW	2.1	Thompson et al. (1994)
Freebayes	0.9.21-g7dd41db	Garrison and Marth (2012)
GapCloser	1.12	Luo et al. (2012)
Gblocks	0.91b	Castresana (2000)
InterproScan	5.20-59.0	Biswas et al. (2002)
Jellyfish	1.1.11	Marçais and Kingsford (2011)
L_RNA_SCAFFOLDER	1.0	Xue et al. (2013)
MAKER	3.0	Holt and Yandell (2011)
PhyML	20120412	Guindon et al. (2010)
prinseq-lite	0.20.3	Schmieder and Edwards (2011)
R	3.3.1	R Core Team (2016)
SAMtools	0.1.19	Li et al. (2009)
SOAPdenovo2	2.04-r240	Luo et al. (2012)
Trimmomatic	0.30 and 0.32	Bolger et al. (2014)
tRNA-scan	1.3.1	Lowe and Eddy (1997)

2.2 *Procambarus* animals

Procambarus animals were obtained from various sources (Table 2.2). *P. virginalis* specimens Petshop, Heidelberg, Moosweiher, and MA1 as well as *P. fallax* Female1 and *P. alleni*

Female1 were provided as previously described (Falckenhayn, 2017; Vogt et al., 2015). In short, strain Petshop was acquired and established in 2004 from a German pet trade (Kölle Zoo). The oldest cultured strain Heidelberg was founded and provided by F. Steuerwald in 2003. *P. virginalis* specimens Moosweiher and Madagascar (MA1) were wild catches from the German lake Moosweiher (located near Freiburg, provided by M. Pfeiffer) and Madagascar (MA1 provided by F. Glaw), respectively. *P. fallax* Female1 (obtained in 2013) and *P. alleni* Female1 (obtained in 2014) were obtained from a German aquarium trade. Additionally, three *P. fallax* animals (Male1 obtained in 2013, Male6 obtained in 2015, and Female4 obtained in 2015) were also acquired from a German aquarium trade. Detailed culturing, tissue dissection, and DNA extraction is described as part of the doctoral thesis of C. Falckenhayn (Falckenhayn, 2017). Sequencing data from the laboratory strain Illinois was provided in a collaboration with W. Stein from Illinois State University (Section 2.3.1).

Madagascar samples

To confirm clonality in distantly evolved populations, five *P. virginalis* specimens were collected from multiples sites in Madagascar (strains MA1-MA5). *P. virginalis* MA1 was collected in Antananarivo and provided by F. Glaw (Vogt et al., 2015; Falckenhayn, 2017). Further Madagascar animals (MA2-MA5) were collected by R. Andriantsoa between March and July 2016 from different regions (MA2: Alaotra-Mangoro, MA3: Analamanga, MA4: Itasy, MA5: Vakinankaratra). For each animal, tissue from the abdominal musculature was extracted and sampled following the protocol described by Falckenhayn (2017).

2.3 High-throughput sequencing

Whole genome sequencing was performed for three different *Procambarus* species, comprising a total of 14 animals obtained from various sources (Table 2.2). Sequencing data was generated in three independent institutions on different Illumina sequencing platforms (Appendix Figure 6.1).

Table 2.2: Origin of *Procambarus* animals. The overview of all *Procambarus* specimens used in this project comprises a total of 14 animals (nine *P. virginalis*, four *P. fallax*, and one *P. alleni*). For each animal, whole genome sequencing was performed.

Species	Specimen	Origin
P. virginalis	Petshop	Petshop laboratory strain
P. virginalis	Heidelberg	Heidelberg laboratory strain
P. virginalis	Moosweiher	wild catch from lake Moosweiher
P. virginalis	Illinois	Illinois laboratory strain
P. virginalis	MA1	wild catch from Antananarivo, Madagascar
P. virginalis	MA2	wild catch from Alaotra-Mangoro, Madagascar
P. virginalis	MA3	wild catch from Analamanga, Madagascar
P. virginalis	MA4	wild catch from Itasy, Madagascar
P. virginalis	MA5	wild catch from Vakinankaratra, Madagascar
P. fallax	Male1	male from aquarium supply
P. fallax	Male6	male from aquarium supply
P. fallax	Female1	female from aquarium supply
P. fallax	Female4	female from aquarium supply
P. alleni	Female1	female from aquarium supply

2.3.1 Genome sequencing

For genome assembly, high coverage whole-genome sequencing was performed for one individual marbled crayfish female from a laboratory strain, herein termed as specimen Petshop. Library generation, sequencing, and preprocessing was performed by Eurofins MWG GmbH (Ebersberg, Germany). Therefore, a total of 100 µg genomic DNA from three different tissues (ovaries, hepatopancreas, and abdominal musculature) was extracted and processed as follows.

For fragmentation of shotgun libraries (SG) a Covaris E210 Instrument was used according to the instructions provided by the manufacturer (Covaris Inc.). Following steps included end-repair, A-tailing, and ligation of indexed adapters. Ligation products were size selected by agarose gels with a targeted insert size of 500 bp. Clusters were generated using an Illumina cBOT and six libraries were sequenced on a HiSeq 2500 platform (HiSeq Control Software 2.0.12.0) with a 2x150 bp paired-end sequencing in rapid mode.

Library generation for long jumping distance sequencing was performed according to

the mate pair library protocol provided by Illumina. The protocol was modified by Eurofins MWG GmbH using adaptor-guided ligation of genomic fragments, which achieves higher accuracies. Targeted insert sizes were 3 kb, 8 kb, 20 kb, and 40 kb. The six LJD libraries were sequenced on an Illumina HiSeq2000 machine with 2x100 bp paired-end sequencing in rapid mode.

Sequencing was run with the original chemistry provided by Illumina (HiSeq Flow Cell v3 and TruSeq SBS Kit v3). Processing of raw output was done using the Real Time Analysis software (RTA, version 1.17.21.3) and CASAVA (1.8.2) to generate and demultiplex FASTQ files (according to the 6 bp index with one mismatch allowed). All reads passed the default Illumina filter. Sequencing quality ($>$31 for SG reads and $>$27 for LJD reads) was estimated by adding a PhiX library before sequencing. Quality values (Q) are provided in Phred quality scores, with the probability of a wrong base call being $P = 10^{\frac{-Q}{10}}$.

Library generation and sequencing of MiSeq reads was performed by the DKFZ Genomics and Proteomics Core Facility (Heidelberg, Germany) following their standard paired end MiSeq protocols (DNA Seq with TruSeq Nano, HiSeq Control Software 2.5.0.5). Raw output was processed using RTA (1.18.54) and FASTQ files were generated using bcl2fastq (1.8.4). Library preparation was done with a targeted insert size of 899 bp and a cycle count of 301.

P. virginalis sequencing data from Illinois

Raw sequencing data from a laboratory strain (Illinois) was acquired as part of a collaboration with Wolfgang Stein from Illinois State University. Therefore, two animals were sequenced as follows. Shotgun libraries were prepared with the Kapa HyperLibrary Preparation kit (Kapa Biosystems). Sequencing was performed on an Illumina HiSeq 2500 machine using TruSeq Rapid SBS kits (version 2.0). FASTQ files were generated and demultiplexed with the bcl2fastq (1.8.4) conversion software. The targeted insert size for SG libraries was 450 bp. Additionally, mate pair libraries with up to 8 kb distance information were generated, though this data was not further used in this project.

Sequencing of additional animals

Library generation of *P. virginalis* specimens Heidelberg, Moosweiher, and MA1 as well as *P. fallax* (Male1 and Female1) and *P. alleni* (Female1) was performed by the DKFZ Genomics and Proteomics Core Facility following the standard procedures as previously described (Vogt et al., 2015; Falckenhayn, 2017). In short, sequencing was done on Illumina HiSeq2000 (V3) machines following the standard Core Facility protocols (R&D protocol for genomic DNA, HCS 2.0.12.0). Raw sequencing output was processed and FASTQ files were generated using RTA (1.17.21.3) and bcl2fastq (1.8.3).

Four *P. virginalis* samples from Madagascar (MA2, MA3, MA4, MA5) and two *P. fallax* samples (Male6 and Female4) were also sequenced by the DKFZ Genomics and Proteomics Core Facility on HiSeq X ultra-high-throughput instruments, following the standard TruSeq DNA protocols (PCR-Free R&D protocols, SBS kit v2.5). Raw sequencing output of Madagascar samples were processed using the RTA base calling software (3.3.76.1) and bcl2fastq (2.18.0.12). FASTQ files for the *P. fallax* samples were generated using bcl2fastq (2.17.1.14).

2.3.2 Genome size estimation

Genome size was estimated *in silico* using raw shotgun reads from the *P. virginalis* specimen Petshop, following the method as described by Li et al. (2010). Therefore, occurrences of short sub strings (k-mers) with length 17 were counted using Jellyfish. The resulting k-mer depths and k-mer frequencies were plotted and peaks were analyzed in R. Then, the average coverage (N) was calculated from the k-mer peak depth (M) as listed below (Formula 2.1).

$$N = \frac{M \cdot L}{L - k + 1} \tag{2.1}$$

where:

N = estimated read coverage
M = k-mer peak coverage depth
L = average read length
k = k-mer size

Eventually, the genome size was estimated by the average coverage and the total number of sequenced bases, which is equivalent to the total sum of nucleotide content in the raw reads (Formula 2.2).

$$G = \frac{T}{N} \tag{2.2}$$

where:

$G =$ Estimated genome size
$T =$ total base pairs

2.4 Genome assembly

Genome assembly was performed in a step wise process based on different algorithms and input data. *De novo* genome assembly comprises steps such as read preprocessing, contig assembly and refinement, scaffolding, and gap filling (Figure 2.1). Applications and parameters for each step are described in the following section.

Genome sequencing was performed as an iterative approach, requiring different methods and parameter settings. Following is the approach for sequencing the most recent version of the *P. virginalis* genome (version 0.4). First, SG and LJD reads were filtered, trimmed, and preprocessed. In detail, raw shotgun (SG) reads were trimmed for adapter sequences and low-quality bases using Trimmomatic (version 0.30). LJD reads were preprocessed with Eurofins software to remove adapter and linker sequences from raw reads. Trimmomatic was used for filtering read pairs (Phred score: 5, window size: 8 bp, Phred score: 15, min length: 30 bp). High quality SG and LJD reads were defined by prinseq-lite. Therefore, reads were filtered for low complexity, gap containments, a GC content below 10% or above 90%, or when they were shorter than 90 bp (SG) and 30 bp (LJD). Similar preprocessing was performed on raw MiSeq data using Trimmomatic (version 0.32) and prinseq-lite.

Contig assembly was performed using the short-read assembler SOAPdenovo2 on shotgun reads with a k-mer size of 75 and sparse pregraph construction. Resources of 10 cores with 20 threads and 200 GB RAM were allocated on the HPC cluster. Sparse pregraph

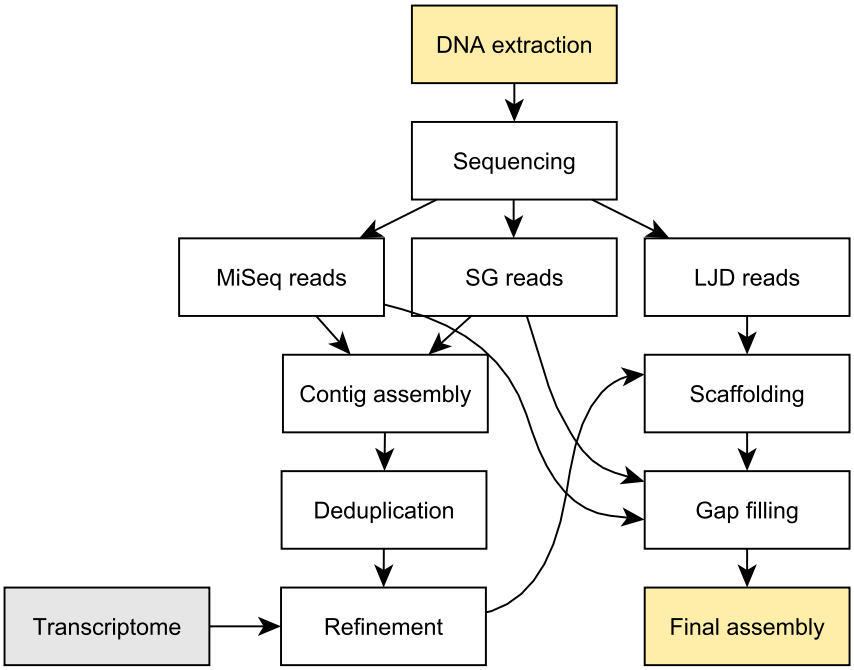

Figure 2.1: Workflow of the marbled crayfish genome assembly. The workflow of the marbled crayfish genome project shortly describes each step for producing the final genome assembly. Incoming edges refer to input data used in the current step whereas outgoing edges are the results. Steps with multiple outgoing edges provide input for more than one successive step. Start and end (DNA extraction and Final assembly) are marked in yellow. External input information is colored in gray.

construction (`sparse_pregraphing` module) allowed for a more time and resource efficient assembly. Extra output information was generated for the assembly process to resolve repeats. Most parameters were set to default apart from omitting kmers (and related edges) with a frequency $\leq$5 (`kmerFreqCutoff` and `kmerEdgeFreqCutoff`) and a predicted genome size of 4 Gbp. Contigs were assembled (`contig` module) with parameters set to resolve repeats, loosely merging similar contigs (`mergeLevel` 3), deleting edges with a coverage $<$5 (`EdgeCovCutoff`), and an arc weight limit of 5 when linearizing two edges within the assembly graph. The output represents the raw contig assembly derived from SG reads.

Contigs were deduplicated because of high heterozygosity and to reduce input complexity for scaffolding. The script `dedupe.sh` from the BBTools was run with parameters for deduplication of identical sequences and containments (or their reverse-complements). Overlaps in sequences were not considered due to the high amount of input sequences.

Furthermore, duplicated sequences were omitted given a minimum identity of 95% and allowing five mismatches (substitutions only) per contig. These values have been chosen based on assuming a high heterozygosity due to a triploid genotype.

After deduplication, transcript-associated contigs were connected by mapping transcriptome information to the contig assembly (transcriptome assembled by Falckenhayn (2017)). Therefore, L_RNA_SCAFFOLDER was used with default parameters. The procedure maps transcripts onto contigs. Eventually, transcript-associated contigs were linked when the required evidence threshold was exceeded.

LJD mate-pairs were used for connecting contigs over long distances, also known as scaffolding. Scaffolding was performed using SOAPdenovo2 (`scaff` module) and a k-mer size of 35 on the HPC cluster with 12 CPUs and 200 GB RAM. After data preparation, LJD reads were mapped onto the contig assembly (default parameters) followed by scaffolding. Parameters were set to use sequences with a minimum length of 100 bp (`minContigLen`), a maximum length difference of 100 bp between estimated and filled gaps (`gapLenDiff`), an upper bound multiplier of 3 for estimating insert sizes (`insertSizeUpperBound`), and using a coverage multiplier of 0.2 for removing contigs within bubbles when coverage is low (`bubbleCoverage`). Weakly connected contigs were kept, due to lower coverage in LJD reads and the complex genome structures.

Four iterations of gap filling, using GapCloser, substantially reduced gap content. Gap filling was performed with a subset of SG reads, a coverage requirement of 6 read pairs, and an average insert size estimate of 450 bp. In total, the subset comprised 26 Gbp in 90 million read pairs (after trimming and quality filtering).

2.4.1 Genome quality

Genome quality was defined by assembly statistics and gene space completeness. Assembly statistics comprise the total number of sequences, total assembly length, and metrics such as the N50 statistic. Gene space was assessed by benchmarking universal single copy arthropod orthologs.

In *de novo* assemblies, regular length measures such as mean and median sequence length is often biased towards very short sequences. Assembly algorithms often fail to incorporate short reads into long consensus structures, resulting in many remaining short

sequences. Hence, statistics such as the mean or median sequence length are skewed, potentially reducing contributions from longer sequences. Thus, a weighted mean was calculated to provide a statistic which gives more importance to longer sequences. The N statistic is a weighted measure for sequence length in an assembly. Therefore, the lengths of sequences within the assembly are extracted and sorted from the longest to the shortest. In an iterative approach, they are summed up (in the direction from the longest to the shortest) and checked if the sum is greater or equal N% of the total assembly length. Provided that is the case, the iteration is stopped and the current length, contributing to the total sum, is the representative N value. The most used parameter for N is 50 (N50). Other frequent parameters are 25, 75, and 90. Higher N values represent an assembly containing longer sequences.

Furthermore, genome quality was estimated using a benchmarking for universal single copy orthologs (BUSCO). BUSCO searches for completed, fragmented, and duplicated orthologs within a given genome. For a conclusive benchmark BUSCO was run on 13 different organisms (including the marbled crayfish) using identical parameters. Organisms were *Nasonia vitripennis* (Jewel wasp, version 2.1, accession GCA_000002325.2), *Locusta migratoria* (Migratory locust, version 1.0, accession GCA_000516895.1), *Stegodyphus mimosarum* (African social velvet spider, version 1.0, accession GCA_000611955.2), *Acanthoscurria geniculata* (Brazilian whiteknee tarantula, version 1.0, accession GCA_000661875.1), *Zootermopsis nevadensis* (Dampwood termite, version 1.0, accession GCA_000696155.1), *Parhyale hawaiensis* (Sand flea, version 3.0, accession GCA_001587735.1), *Ladona fulva* (Scarce chaser, version 1.0, accession GCA_000376725.1), *Bombyx mori* (Domestic silkworm, version ASM15162v1, accession GCA_000151625.1), *Aedes aegypti* (Yellow fever mosquito, version L1, accession GCA_000004015.1), *Drosophila melanogaster* (Fruit fly, version 6.0, accession GCA_000001215.4), *Daphnia pulex* (Common water flea, version 1.0, accession GCA_000187875.1), *Ixodes scapularis* (Black-legged tick, version 1.0, accession GCA_000208615.1), and *Procambarus virginalis* (Marbled crayfish, version 0.4). Parenthesis provide their common name, the assembly version, and a GenBank accession number. BUSCO was run in genome mode using a set of 1066 arthropod orthologs.

2.5 Annotation

For automatic annotation, sequences $\geq$10 kbp were extracted from the assembly. The automatic annotation pipeline MAKER (version 3) was run for eukaryotic organisms. Parameters were mostly set to default. EST evidence was provided by the marbled crayfish transcriptome (Falckenhayn, 2017). Furthermore, protein homology was obtained from the manually curated Uniprot/Swiss-Prot protein database (released on May 2016). Repeat masking was also initiated with default configuration for eukaryotic genomes. A two-times trained Hidden Markov Model was used for automatic gene prediction. Therefore, two runs of automatic annotation were performed using only transcriptome data. Resulting annotation files were used as input to train the Hidden Markov Model of the SNAP gene predictor. Additional gene models from the annotated output of an older BUSCO benchmark (version 1.22) supported automatic annotation. This set contained a total of 203 complete orthologs gene models (out of 2675 orthologs examined by BUSCO version 1.22). Additionally, genes were directly predicted from transcripts and protein homology since *de novo* gene prediction on a fragmented genome can be very strict. Eventually, the EVidenceModeler software combined evidences from *ab initio* gene prediction, homology based alignments, and transcript alignments into consensus gene structures. Each source of automated prediction is clearly comprehensible in the final annotation. This allows to trace the influence of each method within the genome annotation process. Eventually, options to detect alternative splice variants were enabled and predictions were forced for start and stop codons. The maximum intron size was set at 10 kbp and the minimum intron length was set to 20 bp. Single exons were considered (minimum length of 250 bp) as fragmentation of the assembly limits the prediction of genetic structures.

tRNA annotation

Issues in the MAKER configuration prohibited automatic annotation of tRNA isotypes. Thus, tRNAscan-SE was run independently in eukaryotic run mode (tRNAscan and EufindtRNA). tRNAscan parameters were set to strict whereas EufindtRNA parameters were set to relaxed (int cutoff=-32.1). Furthermore, the default covariance model TRNA2-euk.cm was used for prediction. Run statistics and secondary structures were included in the output.

2.5.1 Functional annotation

To characterize protein functions and to further improve annotation, genes and proteins have been assigned functions by homology information. First, all predicted proteins from the MAKER genome annotation have been extracted and blasted (`tblastn` search, E-value cutoff 1e-5) against the annotated marbled crayfish transcriptome (CDS sequences) (Falckenhayn, 2017). Next, proteins were blasted (`blastp` search, E-value cutoff 1e-5) against the Uniprot/Swiss-Prot database (downloaded 10/2016, released on May 2016). The database contains 590,360 protein sequences from a variety of different organisms (including 39,167 different isoforms). Due to more detailed annotations, blast hits from the transcriptome were replaced by Uniprot/Swiss-Prot hits when available. Additionally, functional domains were predicted (InterproScan) and provided as additional evidence for existing gene annotations. This included database cross references (Dbxref) and gene ontology terms (GO).

2.5.2 Pilot analysis on meiosis genes

A set of candidate meiosis genes in the cyclic asexual *Daphnia pulex* was previously described by Schurko et al. (2009). Candidate protein sequences were extracted and searched (`tblastn`, E-value cutoff 1e-5) within an older iteration of the *P. virginalis* genome (version 0.3.2). Potential targets were manually reviewed according to the number of hits, bit score, and E-value.

2.5.3 Web server applications

A central resource on marbled crayfish genome information and data is provided on a web server (http://marmorkrebs.dkfz.de, described in section 2.1.1). The main framework is written in HTML and provides access to web applications. Annotation is integrated and visualized by the Apollo Genome Annotator (version 2.0.2). Apollo enables community driven manual curation and verification of genes. Furthermore, a tool for nucleotide and protein sequence search is integrated in Apollo. Nevertheless, more detailed searches can be done using BLAST. As BLAST was not integrated into Apollo, a custom wwwblast (version 2.2.26) interface was established. There, users can upload nucleotide or protein

sequences and directly search them within the marbled crayfish genome. Enabled routines include `blastn`, `tblastn`, and `tblastx`.

2.6 Genomic analyses

Genomic analyses comprise the phylogenetic relationship based on genomic information of various arthropods, estimation of heterozygosity, and variant analysis of different *Procambarus* species and populations.

2.6.1 Phylogenetic relationships

A set of 138 single copy orthologs (OrthoDB 8) was extracted from recently published arthropod genomes. Genomes include *Acanthoscurria geniculata*, *Aedes aegypti*, *Bombyx mori*, *Daphnia pulex*, *Drosophila melanogaster*, *Ixodes scapularis*, *Locusta migratoria*, *Parhyale hawaiensis*, *Stegodyphus mimosarum*, and *Zootermopsis nevadensis*. Accession numbers and assembly versions were already described in Section 2.4.1. Additionally, genomes of *Apis mellifera* (Western honey bee, version 4.5, accession GCA_000002195.1) and *Camponotus floridanus* (Florida carpenter ant, version 3.3, accession via http://hymenopteragenome.org/camponotus/?q=genome_consortium_datasets) were used in this analysis. To estimate genetic divergence, a multiple sequence alignment has been calculated with ClustalW (default parameters). Alignments were reduced to focus conserved blocks using Gblocks. Eventually, a phylogenetic tree was constructed based on maximum likelihood estimation using PhyML (default parameters).

2.6.2 Variation analysis

Shotgun reads of all *Procambarus* individuals (*P. virginalis*, *P. fallax*, and *P. alleni*) were mapped to the marbled crayfish reference genome using Bowtie 2. Subsequently, duplicates were removed and a subset of sequences $\geq$10 kb was extracted with SAMtools. A total of 665,574,093 callable sites, comprising about 19.0% of the total assembly size, were identified by excluding gap bases and remapping Petshop reads. Variant calling was performed using Freebayes with parameters set for a detection probability of at least 70%, a

minimum mapping quality of 30, a minimum base quality of 20, and a coverage of at least 6 reads per site. Furthermore, all haplotypes were reported and additional variant evidence was provided by the reference sequence. Data for multiple *P. virginalis* and *P. fallax* animals allowed for a multi-sample variant calling, while variants for *P. alleni* were called on one individual sample. Furthermore, *P. virginalis* samples were analyzed for a triploid genotype, while the other two species were analyzed in diploid mode. Variants were filtered for type (SNPs) and a quality score ($\geq$100).

Comparative analyses include all *Procambarus* individuals, except the four high-coverage Madagascar samples MA2, MA3, MA4, and MA5. Variant calls for *P. virginalis* specimen Petshop (remapped genome assembly reads) were used to filter potential technical errors. Therefore, valid sites were defined by at least one reference read observation in the genome individual (Petshop). Additionally, all sites required a valid genotype in all samples, as defined by Freebayes. Filtering and processing was performed using a custom script and the statistical computing package R.

Polymorphic sites were extracted to create variant relationships. First, heterozygous sites in the reference individual (Petshop) were eliminated. This enabled a reduction of sites to strictly non-heterozygous loci. Polymorphic sites in marbled crayfish samples were identified by at most one single base nucleotide substitution to the reference sequence. For the diploid *P. fallax* and *P. alleni* individuals, all types of variants (homozygous/heterozygous variants) were considered. Sequence similarity was estimated by pairwise comparison of biallelic variant loci. Therefore, the number of identical variation sites was calculated for each sample pair. Counts were normalized for the total variant number in each sample. Next, a maximum distance matrix was calculated and clustered hierarchically using the complete linkage method. Finally, an unrooted phylogenetic tree was generated using the APE package in R (Paradis et al., 2004).

Biallelic variant distribution was defined by the ratio of alternative read observations to the total observations per site, e.g. a ratio of 0.5 describes 50% alternative and 50% reference reads in one specific site. Allele distributions of heterozygous and homozygous loci were illustrated in violin plots using the R package vioplot. Variant distributions were combined by species.

Differentiation of biallelic and triallelic genotypes was defined by the number of different

alleles. Genotypes with at most two distinct alleles to the reference genome are defined as biallelic variants whereas genotypes with three distinct alleles are defined as triallelic variants. Again, homozygous and heterozygous variants were considered for each sample. The ratio of biallelic to triallelic variants was calculated and visualized in R.

Circos plot of Madagascar samples

Analysis of polymorphic sites in distant populations included variant data from *P. virginalis* specimens Petshop, MA1, MA2, MA3, MA4, and MA5. Furthermore, variant information from one *P. fallax* (Female4) was used for comparison. Eight random sequences were selected from a subset of high nucleotide density genome sequences. Variant sites in these sequences were filtered for single base substitutions as described above. Eventually, sites were plotted using the OmicsCircos package in R (Hu et al., 2014).

2.6.3 Heterozygosity levels

High coverage sequence information of *P. virginalis* specimen Petshop was used to estimate heterozygosity levels. Therefore, shotgun reads were remapped onto the genome assembly (limited to sequences $\geq$10 kbp) as previously described (Section 2.6.2). Variant calling was performed using Freebayes with identical settings as described above. Heterozygous positions were extracted after filtering a variant quality of at least 30 and nucleotide coverage of at least 15. Heterozygosity information for other genomes, estimated by similar approaches, were extracted from publications (Sanggaard et al., 2014; Wang et al., 2014; Albertin et al., 2015). To estimate heterozygosity levels in *P. fallax*, a preliminary raw contig assembly was generated. Therefore, high coverage sequencing data from *P. fallax* Female4 was assembled into contigs using SOAPdenovo2 (similar parameter settings as already described in section 2.4). Variant calling was performed similar as in the marbled crayfish. The only exception was a diploid ploidy level and, due to lower sequencing coverage, a depth cutoff of 10. Heterozygosity levels were calculated as the total number of heterozygous sites divided by the number of total sites (disregarding ambiguous gap bases).

2.6.4 Single base substitutions and mutation signatures

To estimate single base substitutions in marbled crayfish, all sequenced *P. virginalis* animals were considered. Filters for variant sites were applied as previously described (Keightley et al., 2009, 2015). Remapping the reference genome reads (Petshop) allowed for filtering sites prone to technical errors. Therefore, valid sites were defined to have reference read coverage of $\geq$10 and $\leq$200. Furthermore, sites were only considered when the reference animal did not have any alternative allele observations. All samples were required to have a valid genotype for each site, i.e. each allele is clearly defined by Freebayes. A mutation is allowed to occur in at most two independent samples. Eventually, single nucleotide substitutions are defined as homozygous substitutions to the reference genome.

Substitutions provide mutation signatures by analyzing transitions from the reference base to the substituted base. Signatures for each sample were combined and plotted in R. The AT:CG ratio was calculated as the ratio of C/G $\rightarrow$ A/T counts to A/T $\rightarrow$ C/G counts (Keith et al., 2016). The Ts:Tv ratio results from the ratio of transitions (A $\leftrightarrow$ G and T $\leftrightarrow$ C) to transversions (A/G $\leftrightarrow$ C/T) (Keith et al., 2016).

3 Results

Results comprise the assembly of sequencing data, obtained from one individual female marbled crayfish, into a first *de novo* draft assembly. This includes steps such as contig assembly, refinement using the transcriptome, and connecting sequences using long jumping distance information. Moreover, genetic features were predicted and mapped onto the novel assembly. The marbled crayfish genome provides detailed insights into genomic characteristics and allows for comparative analyses on sequence variation.

3.1 Sequencing

Rapid development of sequencing technologies enables highly accurate base calling of DNA. Here, Illumina HiSeq machines were used for a cost efficient high coverage short read sequencing of genomic data. Additional *Procambarus* species were sequenced for downstream analyses, e.g. to characterize the evolution of the parthenogenetic marbled crayfish.

3.1.1 Genome sequencing

Whole genome sequencing was performed on a single individual female marbled crayfish obtained from a pet trade, herein denoted as specimen Petshop. Three different tissues, namely ovaries, hepatopancreas, and abdominal musculature, were used for library generation. Different sequencing approaches were pursued due to the large genome size and to bypass limitations of short read sequencing.

Shotgun sequencing (SG) was mainly used to produce continuous consensus sequences (contigs). High coverage data substantially decreases sequencing errors. Six

lanes of SG data were sequenced on an Illumina HiSeq 2500 machine with a targeted insert size of 500 bp, producing a total of 245,157 Mbp raw yield in 817 million read pairs (Table 3.1). Furthermore, an additional MiSeq library was produced with the initial purpose to extend and improve the contig assembly provided by Eurofins MWG GmbH (Ebersberg, Germany). However, producing a new contig assembly allowed for an a priori integration of these reads. Sequencing on the Illumina MiSeq platform yielded 11,218 Mbp raw information in 81 million read pairs (Table 3.1).

Six lanes of long jumping distance (LJD) reads were generated to connect contigs into long scaffolds. Long jumping distance reads were obtained by sequencing modified vectors of known size, which resulted in mate pairs separated by 3 kbp, 8 kbp, 20 kbp, and 40 kbp. Sequencing on an Illumina HiSeq 2000 machine produced a total of 94,077 Mbp raw yield. In particular, targeted insert sizes of 3 kbp yielded 18,005 Mbp raw information in 90 million read pairs, 8 kbp yielded 33,866 Mbp in 169 million, 20 kbp yielded 25,950 Mbp in 130 million, and 40 kbp yielded 16,256 Mbp in 81 million read pairs (Table 3.1).

In total, a raw yield of 350,452 Mbp was produced from whole genome sequencing on different technologies (Table 3.1). This corresponds to a total genome coverage of about 100x when assuming a 3.5 Gbp large genome.

Table 3.1: Genome sequencing results. Raw whole genome sequencing results from one individual female. Data was generated on different Illumina sequencing platforms. High coverage shotgun sequencing was produced in paired-end mode and a targeted insert size of 500 bp. Long jumping distance libraries (LJD) were sequenced in a mate pair approach with targeted insert sizes 3 kbp, 8 kbp, 20 kbp, and 40 kbp. Read pairs are provided in millions ($\times 10^6$) and raw yield in Mbp (10^6 bp).

Type	Insert size (bp)	Read pairs ($\times 10^6$)	Yield (Mbp)	Protocol
Shotgun	500	817	245,157	HiSeq 2500 PE150
Shotgun	550	19	11,218	MiSeq PE300
LJD	3,000	90	18,005	HiSeq 2000 PE100
LJD	8,000	169	33,866	HiSeq 2000 PE100
LJD	20,000	130	25,950	HiSeq 2000 PE100
LJD	40,000	81	16,256	HiSeq 2000 PE100
Total		1,306	350,452	

Genome size estimation

Whole genome sequencing data allows for computational insights into genome structure, e.g. by analysis on nucleotide distribution patterns. Counting sub strings of fixed length not only provides preliminary insights into genomic characteristics, such as repeat content and heterozygosity, but also allows for *in silico* genome size estimation. Therefore, sub strings of size 17 (17-mers) were counted from raw SG data of *P. virginalis* Petshop. When plotting the distribution of k-mer frequencies, two distinct peaks were observed at k-mer depths of 16 and 45 with frequencies of 17,527,603 and 15,706,869 respectively (Figure 3.1). The first peak, at k-mer depth 16, represents heterozygous k-mers whereas the second peak represents homozygous k-mers. First insights into genome structure can be derived from the heterozygous peak occurring at a third of the depth ($\frac{n}{3}$) as the homozygous peak (n). This strongly correlates with the triploid genotype of marbled crayfish (Vogt et al., 2015; Martin et al., 2016). High k-mer frequencies at depths $<$5 mainly occur due to technical artifacts.

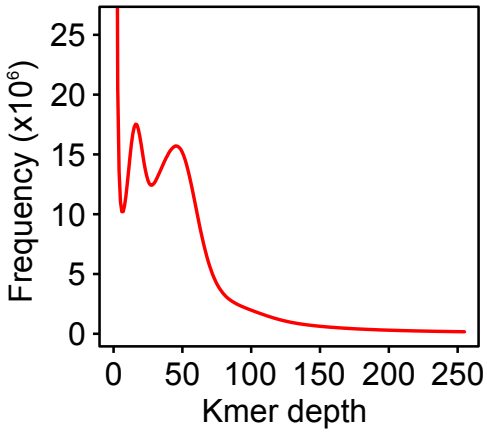

Figure 3.1: K-mer frequency and depth plot. The frequency and depth of k-mers ($k = 17$), derived from shotgun sequencing reads, results in two distinct peaks at depth 16 (with a frequency of 17,527,603) and 45 (with a frequency of 15,706,869). K-mer frequencies for a depth $<$5 are cut as they represent technical artifacts. Frequencies are provided in millions ($\times 10^6$).

In silico genome size estimation can be done as previously described (Section 2.3.2). In combination with results from flow cytometry the haploid genome size was estimated at 3.5 Gbp (Falckenhayn, 2017) .

3.2 Assembly

The ultimate goal of genome assembly is to build the longest possible consensus sequences to represent individual chromosomes. Therefore, short reads were assembled by constructing de Brujin graphs as implemented in SOAPdenovo2.

A first raw contig set was produced by trimming and filtering SG and MiSeq reads. Multiple contig sets were generated using different assembly parameters (data not shown). Empirically, the contig set with the fewest sequences, highest nucleotide content, highest mean sequence length, and highest N50 value was selected. The resulting contig set consists of almost 19 million contigs, 3.2 Gbp total bases, and a GC content of 43.5%. The shortest sequence resulted from an unassembled read with 76 bp whereas the longest consensus sequence reached over 17 kbp. Average sequences length is 170 bp, closely correlated with the N50 of 197 bp.

Owing to a triploid genotype and high heterozygosity, contig assembly contained a high number of identical or highly similar sequences. This substantial increase in input complexity is a potential problem for subsequent assembly steps. Deduplication filtered highly identical sequences ($\geq$ 97%), drastically reducing the assembly set by 72.6% (to 5.2 million sequences). Consequently, the number of total bases drops by 45.9% (to 1.75 Gbp). The average length of filtered sequences (108 bp) suggests that most information loss occurs in sequences which could not be integrated in the contig assembly and, due to their size, are not usable for further processing steps.

Prior analyses on genome completeness showed fragmentation and absence of important genetic structures. To produce a high-quality genome assembly, contig refinement was required. Transcriptome information (provided by C. Falckenhayn) was used to correct protein coding structures. Transcripts were used to orientate and combine contigs coding for the same genes. The initial sequence set could be reduced by 60,151 sequences (1.1%) while total sequence length increased by 6.2 Mbp. The increase in sequence length is due to the introduction of gaps when two contigs match onto a transcript but are not overlapping.

Most contigs were limited by size due to complex genome structures. Unresolved complexities cause fragmented sequences in the assembly algorithm. Several approaches exist for improving assembly structure by combining short contigs into longer sequences (scaf-

folding). Here, long jumping distance reads were generated from three different tissues, like in shotgun sequencing. Long jumping distance reads are produced by incorporating mate pairs into vectors of known length. This technique allowed to obtain distances from 3 kbp, 8 kbp, 20 kbp, and 40 kbp. SOAPdenovo2 provides a scaffolding module which allowed to produce sequences (also known as scaffolds) of up to 718 kbp. Contigs were connected such that the weighted mean sequence length (N50) increased to 28,228 bp.

Nevertheless, scaffolding also introduces gaps in the genome assembly. Gaps occur when only the distance (not the nucleotides) between two contigs is known. In other terms, two contigs are connected by mate pairs but neither are they overlapping nor any nucleotide information in the connection is known. Still, the connection provides useful information for later analyses. Consequently, the unknown sequence within a known distance is filled with gap bases (nucleotide N). Gaps can span a few hundred to several thousand nucleotides. Further refinement in *de novo* genome assembly aims to decrease the total number of gaps and gap bases. Gap refinement eliminated a total of 595,696 gaps (37.7%) in four iterations (Table 3.2). This reduced the number of total gap bases by 161,096,220 bp (Table 3.2).

Table 3.2: Gap filling results. Provided is the total number of gaps and gap bases before each iteration of gap filling. Gap filling was performed in four iterations. Extended and Finished describes the total numbers of gap bases and gaps which could be reduced after each iteration of refinement.

Iteration	Gap sum	Extended	Gap count	Finished
1	1,814,626,668	140,655,387	1,579,001	558,450
2	1,680,463,365	16,235,535	1,020,551	31,297
3	1,665,630,287	3,147,864	989,254	4,147
4	1,663,096,159	1,057,434	985,107	1,802
Total		161,096,220		595,696

Eventually, the final assembly contains 3.4 million sequences (contigs and scaffolds), comprising about 3.5 Gbp of bases (Table 3.3). The shortest sequence length is 100 bp and the longest is almost 718 kbp. Mean and median sequence length are 1,034 bp and 226 bp, respectively. The weighted mean statistic N50 is 28,228 bp and N90 is at 255 bp. Nucleotides are distributed by 532,071,537 Adenines (15.2%), 403,641,751 Cytosines (11.5%), 516,051,095 Thymines (14.7%), 397,229,790 Guanines (11.3%). Almost half of

the assembly (47.3%) consists of gap bases (1,662,662,583) which distribute over 983,305 regions (Figure 3.2A).

Table 3.3: Genome assembly statistics. Various assembly statistics for the final genome assembly. The full assembly contains all sequences obtained from genome assembly. Statistics for sequences ≥1 kbp and ≥10 kbp are provided as they were used for downstream analyses.

Description	Full Assembly	Assembly≥1 kbp	Assembly≥10 kbp
Total entries	3,394,710	240,964	52,418
Total length	3,511,656,756	2,664,252,624	2,226,816,413
Min length	100	1,000	10,000
Max length	717,999	717,999	717,999
Mean length	1,034	11,056	42,481
Median length	226	1,723	35,822
N50	28,228	41,510	48,916
N90	255	4,509	20,653
GC content	43.31%	43.25%	43.23%
Gap bases	1,662,662,583	1,659,644,686	1,549,212,759
Number of gaps	983,305	782,897	497,832

A filter for minimum sequence length is often applied when processing with large genome assemblies. While sequence information gets lost, assembly complexity is substantially reduced. Most applications benefit from a lower input complexity, especially when short sequences play no important role, e.g. as for automatic annotation or comparative analyses. Short sequences mostly remain from not being integrated into a consensus structure due to technical artifacts (sequencing errors) or biological variations (heterozygous positions). Nevertheless, enforcing a minimum sequence length reduces the nucleotide content within the assembly at a substantially slower rate than the number of sequences (Figure 3.2B). This justifies a reduction of input complexity for further analyses without losing much sequence information.

3.2.1 Genome assembly quality assessment

Assembly quality tries to measure the completeness of an assembly. A high-quality genome assembly allows for more confident results in downstream analyses. For the marbled crayfish genome, the quality was measured by sequence statistics and gene completeness.

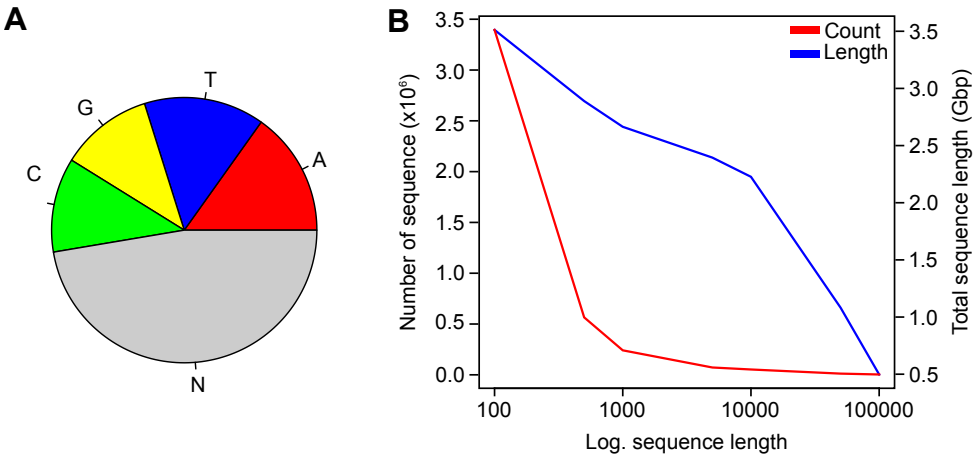

Figure 3.2: Nucleotide distribution and cumulative sequence lengths. Genome assembly statistics are provided by the nucleotide distribution and an overview of cumulative sequence lengths. (A) Total nucleotide distributions within the assembly: Adenine (A, red) with 15.2%, Thymine (T, blue) with 14.7%, Guanine (G, yellow) with 11.3%, and Cytosine (C, green) with 11.5%. Unknown gap bases are denoted with the nucleotide N (gray). Gap bases are most abundant with a proportion of 47.3%. (B) Cumulative sequence lengths illustrate the correlation of nucleotide content (in Gbp, colored in blue) and the number of sequences (in millions, colored in red). Upon enforcing a minimum length cutoff (logarithmic), the reduction of sequences is substantially higher than the reduction of nucleotide content.

Statistics such as the total assembly length, average sequence length, number of scaffolds, and base distribution provide initial estimations on the assembly set. Nevertheless, mean sequence length is often skewed since assemblies are plagued by short contigs ($\leq$ average read size) which resulted from not being built into a consensus structure. Especially, in assemblies of high heterozygous or repeat rich genomes a large proportion of short sequences remain. Therefore, a weighted mean is adapted to reduce the impact of short sequence on the mean sequence length. The N statistic is a measure where sequences of equal length or longer are representing at least N% of the total assembly size (more details in the methods section). The most common parameter for the N statistic is 50, also known as the N50. Often additional parameters such as N25, N75, or N90 are provided. The N50 of the marbled crayfish genome is 28,228 bp while the N90 value is 255 bp, indicating a high amount of short sequences in the assembly (Figure 3.3A).

The benchmarking tool BUSCO assesses the representation of universal single copy orthologs within a given genome. BUSCO was used to compare the marbled crayfish as-

sembly with other recently published arthropod genomes. From a set of 1066 arthropod orthologs, 87.8% were found complete and 7.4% were found fragmented (Figure 3.3B). Fragmented gene representation is defined by finding only parts of a gene. About 1.1% of orthologs were found duplicated and 4.8% were entirely missing. This ranks the marbled crayfish assembly of the marbled crayfish closely to other recently published arthropod genomes (rank 9 of 13, sorted by complete gene structure) (Figure 3.3B).

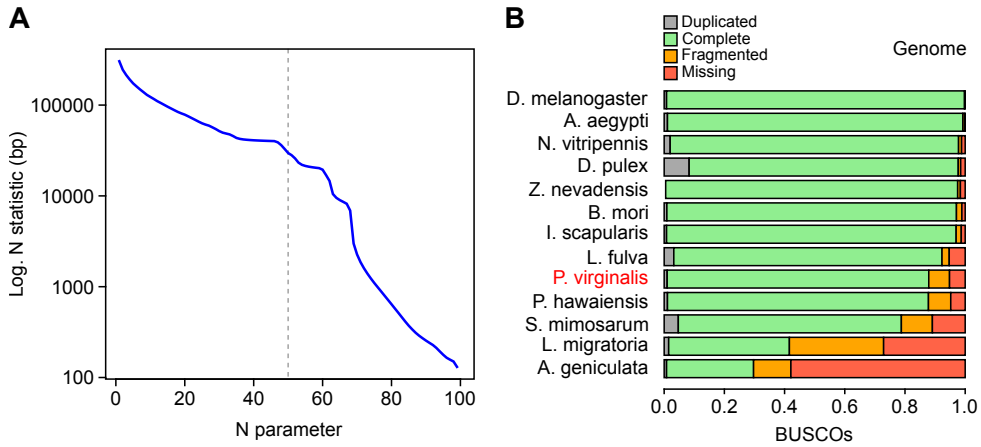

Figure 3.3: Assessment of assembly quality. Genome quality was assessed by two approaches. (A) The N statistic (logarithmic) provides a measure for the weighted mean sequence length. Common parameters are the N50 (dashed gray line, 28,228 bp) and N90 (255 bp). (B) BUSCO analysis of 1,066 universal single copy orthologs in 13 different arthropods species. The *P. virginalis* genome assembly ranks in mid field with 87.8% complete (1.1% duplications), 7.4% fragmented, and 4.8% missing genes. Complete genes are colored green (duplications in gray), fragmented are colored orange, and missing fractions are colored red.

Phylogenetic analysis based on gene models

Conserved domains and protein structures provide useful information for analyzing evolution and relationships between different species. For phylogenetic placement, a set of 138 single copy orthologs from the arthropod phylum was extracted and analyzed. These genes are highly conserved across different arthropods and provide insights into evolution. Clustering the marbled crayfish next to *P. hawaiensis* and near the water flea *D. pulex* confirmed placement within the crustacean lineage. Due to the lack of genomic information of closer related species a more detailed analysis was not possible.

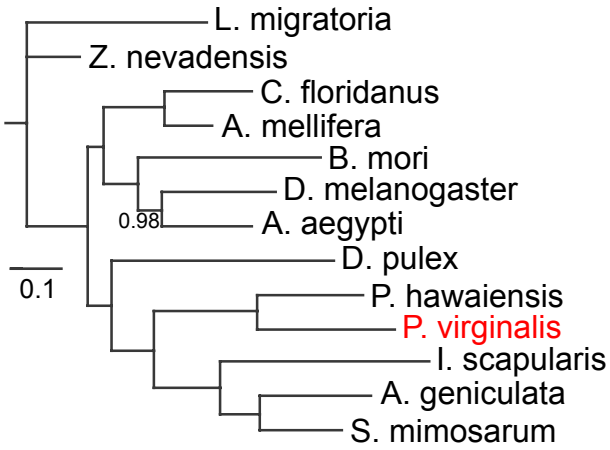

Figure 3.4: Phylogenetic placement of *P. virginalis* based on protein homology. Phylogenetic placement of *P. virginalis* was performed by analyzing 138 universal single copy orthologs among arthropod species. Orthologs were reduced to conserved blocks and the phylogenetic tree was generated based on maximum likelihood estimation (Scale bar: 0.1). Bootstrap confidences $<$1.0 are provided at the respective branches.

3.2.2 Transcriptome assembly of the noble crayfish, *A. astacus*

To further elucidate gene space in freshwater crayfish the transcriptome of the noble crayfish *Astacus astacus* was sequenced and assembled by Theissinger et al. (2016). Personal contributions to the study comprised the analysis of transcriptome completeness. The benchmarking pipeline as well as additional resources were already established for the marbled crayfish project (BUSCO version 1.22). From a total of 2765 universal single copy arthropod orthologs, 64% were found complete, 4.8% were found fragmented, and 30% were entirely missing.

3.3 Annotation

The automatic annotation pipeline MAKER provides methods for homology based annotation and *ab initio* prediction. Due to a high occurrence of short sequences, annotation was limited to sequence $\geq$10 kbp. Genomic evidence of shorter sequences is often considered not to contain meaningful information. Homology models were annotated using the Uniprot/Swiss-Prot database, comprising 553,941 manually verified transcripts from various

organisms (June 2015). Further input for gene model prediction was obtained from BUSCO annotations (version 1.22) and RNA-seq evidence from the marbled crayfish transcriptome (Falckenhayn, 2017).

The MAKER annotation pipeline could predict 21,772 gene models with 22,205 transcripts, including alternative splice variants. Genes incorporate a total of 86,771 exons and 65,683 introns. Total exon length is 22 Mbp (253 bp on average) and total intron length is 129 Mbp (1960 bp on average) (Figure 3.5). On average, genes and transcripts span 6.7 kbp with the longest gene being up to 100 kbp (Figure 3.5). This ranks the marbled crayfish in between average genetic feature lengths of *Daphnia pulex* (intron: 0.3 kbp, gene: 2 kbp) and *Parhyale hawaiensis* (intron: 5.4 kbp, gene: 20 kbp) (Colbourne et al., 2011; Kao et al., 2016).

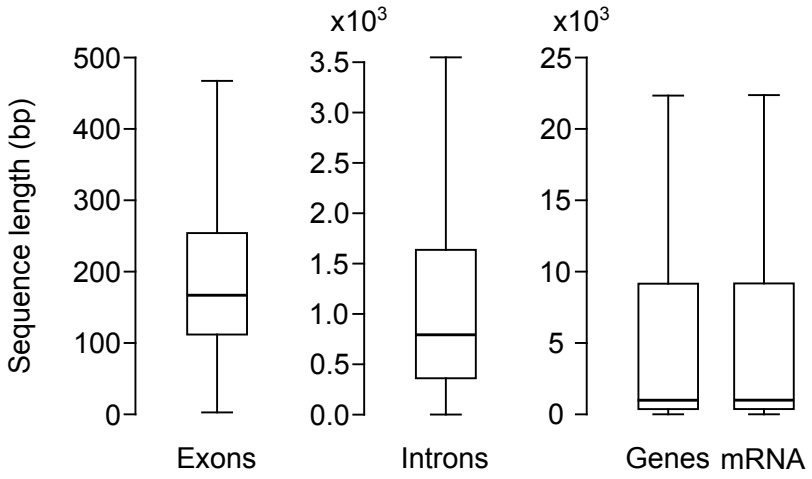

Figure 3.5: Genomic features length distribution. Genomic features were automatically annotated using MAKER. Boxplots illustrate genomic feature length of exons, introns, genes, and transcripts. Lengths of introns, genes, and transcripts are provided in kbp ($\times 10^3$ bp).

Repeat detection

The MAKER pipeline combines two algorithms for automated repeat detection. RepeatMasker masks DNA sequences for features such as SINES, LINES, and low complexity regions. Therefore, it uses alignments of a manually reviewed nucleotide repeat library. Repeatrunner masks repeats from a database such that it identifies protein coding portions of retro-elements and retro-viruses.

In summary, *de novo* repeat detection predicted a total of 481,851 repeats, comprising about 8.8% of the genome. Repeats are highly abundant for simple repeats (54.3%) and LINEs (15.6%). SINEs and DNA transposons comprise each of about 8.4%, followed by low complexity regions (5.7%), LTRs (5.2%), and some other lowly abundant repeat classes (combined proportion of 2.4%) (Figure 3.6).

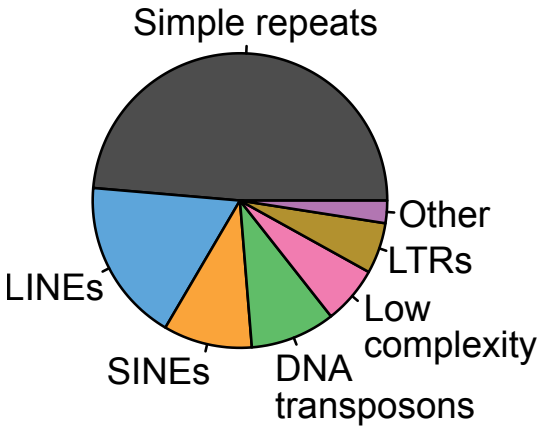

Figure 3.6: Proportions of annotated repeat classes. Automatically annotated repeats were classified into seven major classes. Proportions are shown counterclockwise in decreasing order. Simple repeats being most abundant with 54.3% (gray), followed by LINEs with 15.6% (blue), SINEs with 8.4% (orange), and DNA transposons 8.4% (green). Low complexity regions make up 5.7% (pink) and LTRs 5.2% (brown). The combined proportion of lowly abundant repeat classes (Other) was 2.4% (purple).

3.3.1 tRNA annotation

Transfer RNA (tRNA) genes were automatically annotated to further characterize genomic content. In short, probabilistic profiles of secondary structure were searched within the assembly. Prediction found a total of 17,990 possible tRNAs of which 16,596 are assumed to be pseudogenes. 1,356 genes were found to decode for the standard 20 amino acids. Most abundant tRNA isotypes were Threonine (666) and Valine (226) (Figure 3.7). Least abundant were the isoforms Histidine (5) and Tryptophan (4).

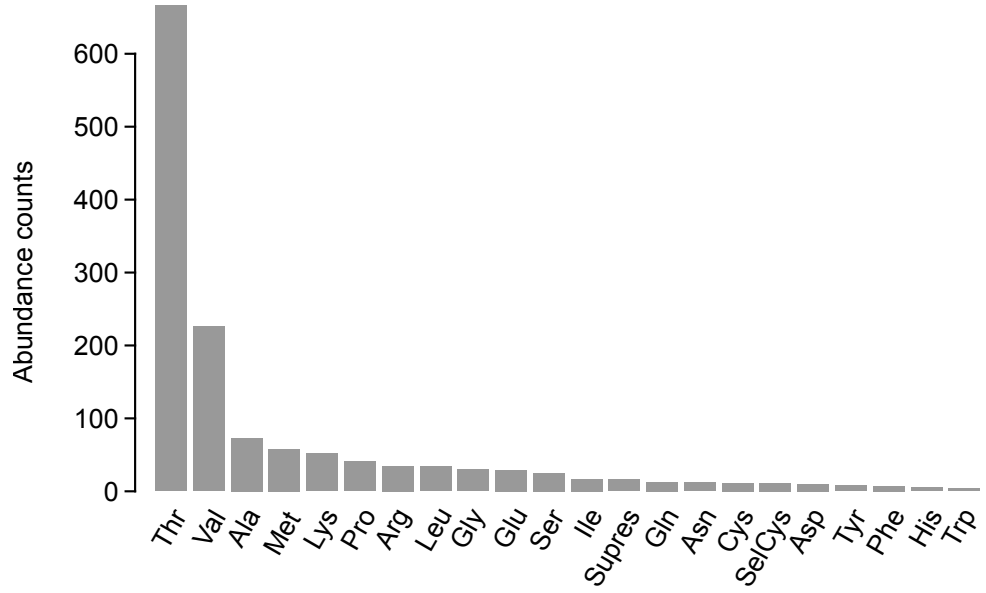

Figure 3.7: Abundance of predicted tRNA isotypes. Abundance of tRNA isotypes were predicted using tRNAscan-SE. Isotypes are provided in 3-letter code. The Supres label represents tRNA suppressors coding for stop codons.

3.3.2 GH9 superfamily – annotation par example

To further elucidate assembly and annotation quality, the gene structure of one particular gene was investigated. Cellulase activity in crayfish has been previously described to potentially play a key role in omnivorousness (Byrne et al., 1999; Crawford et al., 2005). Thus, a cellulase gene from the glycoside hydrolase family 9 (beta-1,4-glucanase) was searched within the marbled crayfish genome assembly (CAZy database). Ultimately, evidence for at least one highly conserved GH9 gene was found, composed of 13 coding exons (Figure 3.8). Regarding sequence length, the CDS comprised 2 kbp and was encoded within a gene of about 9 kbp.

3.3.3 Preliminary analysis on meiosis genes

Apomictic parthenogenesis as reproduction mode in marbled crayfish implies the aberrations in the process of meiosis (Scholtz et al., 2003; Vogt et al., 2004). This ultimately led to the investigation on meiosis relevant genes in the marbled crayfish genome. Parthenogene-

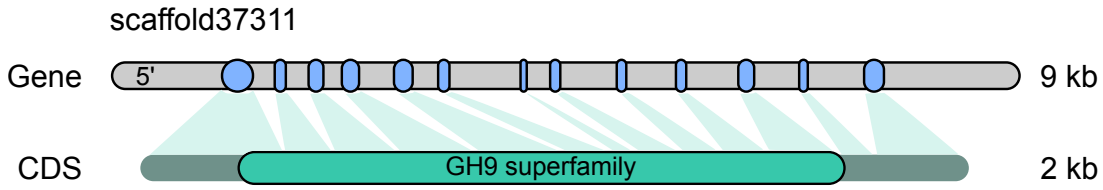

Figure 3.8: GH9 superfamily annotation example. Automatic annotation revealed gene structures such as a gene from the GH9 superfamily important for cellulose digestion. The schematic of the gene shows the CDS (derived from the transcriptome) and the assembly region (scaffold37311), containing the genes intron (gray) and exon (light blue) structure. Gene length is about 9 kbp whereas the CDS is only about 2 kbp.

sis and meiosis related genes have already been described in the water flea, *Daphnia pulex* (Schurko et al., 2009). A set of 93 meiosis relevant genes were extracted and searched within an older version of the genome (0.3.2). Interestingly, most meiosis relevant genes were either found complete or fragmented (n=84). Nevertheless, evidence was lacking for genes such as Rec8 and Msh5.

3.3.4 Functional annotation

Automatic annotation in *de novo* genome projects often predicts genes without a direct association of their function. Evidence from homologous proteins already exist (MAKER annotation) but were not clearly defined for predicted genes. Explicit gene associations are required for pathway and comparative analyses. Hence, functional annotation was performed by extracting predicted protein sequences and realigning them onto the manually verified Uniprot/Swiss-Prot database. In total, 66.8% (of 21,773) predicted genes were mapped to a protein of known function. Annotations from transcripts of the marbled crayfish transcriptome were provided as additional information.

Furthermore, protein signatures were searched using InterProScan to obtain functional domains. InterProScan provides a large resource of predictive models derived from several different databases. Models were carefully picked and curated by the InterPro consortium. Eventually, 41.6% genes were assigned with functional domains from InterProScan. Moreover, 129 unknown genes could be annotated with domains. The InterProScan pipeline annotated a total of 10,251 genes with database cross references (Dbxref) and 6,457 genes with gene ontology terms.

In total, 67.4% of genes were annotated with a functional annotation either by complete protein evidence or functional domain homology. This provides a useful addition for the *de novo* annotation of the marbled crayfish genome assembly.

3.3.5 Manual curation on a designated web server

Automatic annotation was created to be a powerful support for experimental annotation procedures. Manual refinement is reasonable and advisable as algorithmic annotation is prone to errors. A web server was established providing a central resource for manual curation and genomic analysis of the marbled crayfish genome (http://marmorkrebs.dkfz.de). The website comprises useful information about the marbled crayfish project as well as powerful tools like the Apollo genome annotation editor and a BLAST search. The browser application Apollo allows users to view and edit predicted genome annotation. Annotations are provided with distinct identifiers and their respective algorithmic sources (Figure 3.9). Moreover, curators have the possibility of editing and refining annotation in regions of interest. Every change in annotations is traceable and can be complemented by comments or references.

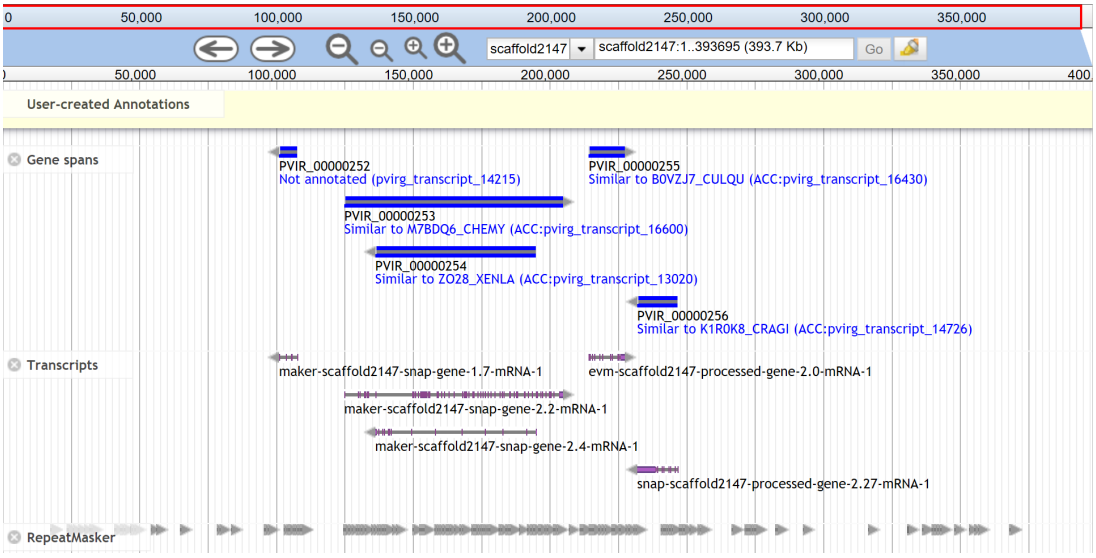

Figure 3.9: Apollo browser example for genome annotation. The Apollo genome annotation browser and editor was established as a community driven resource for the marbled crayfish genome annotation. Annotations are provided with the source of each evidence.

3.4 Variation analysis

Providing a genome assembly allows comparative analyses of sequence variation profiles. Sequencing data from 14 different animals was collected: nine *P. virginalis*, four *P. fallax*, and one *P. alleni*. Then, analyses on sequence similarity, polymorphisms and heterozygosity were performed.

3.4.1 Mapping coverages from *Procambarus* samples

Sequencing of additional animals was done to compare genomic data of different origins (Table 2.2). Therefore, several approaches were pursued resulting in varying sequencing yields (Appendix Table 6.2). Genome coverage was calculated after trimming, filtering, and mapping of reads (Table 3.4). Remapping reads, used for genome assembly (*P. virginalis* Petshop), resulted in the highest overall coverage of 71.8x. As *P. virginalis* Heidelberg was previously intended to be used for genome sequencing high coverage sequencing data was generated (54.3x). Individuals sequenced on single lanes (*P. virginalis* specimens Moosweiher, MA1, Illinois; *P. fallax* Male1, Female1; *P. alleni* Female1) produced coverages of 16.2x to 20x. Higher yields were obtained for individuals sequenced on Illumina HiSeq X ultra-high-throughput machines, resulting in coverages from 27.6x (*P. virginalis* MA5) to 41.6x (*P. fallax* Male6).

3.4.2 Genetic variant calling

For each animal, genetic variants were called using Freebayes. To increase sensitivity of variant detection, data from *P. virginalis* and *P. fallax* was processed in a multi-sample approach. Single sample calls were done for *P. alleni*, since only one individual was available. Variant calls were considered in 665,574,093 genomic sites, comprising about 19% of the total assembly size.

3.4.3 Heterozygosity rate

Estimating the heterozygosity is an important step for understanding the structure and complexity of a genome. In the marbled crayfish, heterozygosity provides insights into evolu-

Table 3.4: Mapping coverages for *Procambarus* animals. Total genome coverage of SG sequencing data after trimming, filtering, and mapping. The coverage was calculated as the sum of coverages per nucleotide divided by the total number of nucleotides (after mapping).

Species	Specimen	Coverage
P. virginalis	Petshop	71.8x
P. virginalis	Heidelberg	54.3x
P. virginalis	Moosweiher	15.6x
P. virginalis	Illinois	16.2x
P. virginalis	MA1	16.8x
P. virginalis	MA2	30.0x
P. virginalis	MA3	36.0x
P. virginalis	MA4	29.1x
P. virginalis	MA5	27.6x
P. fallax	Male1	19.9x
P. fallax	Male6	41.6x
P. fallax	Female1	17.8x
P. fallax	Female4	38.7x
P. alleni	Female1	16.6x

tion and origination of the species. To estimate heterozygosity in *P. virginalis*, all valid loci were extracted from variant calling analysis and the number of heterozygous positions in the analyzed sequence set (excluding gap sequence) was calculated. This resulted in a genome wide rate of 0.53% (Figure 3.10). In *P. fallax*, estimation was performed in a similar approach. Prior to analysis a preliminary contig assembly was produced using SOAPdenovo2. Subsequently, *P. fallax* reads were remapped onto the preliminary assembly and the heterozygosity rate was estimated at 0.03% (Figure 3.10). Previous publications on genome assemblies provided heterozygosity rates calculated by similar approaches (Figure 3.10) (Sanggaard et al., 2014; Wang et al., 2014; Albertin et al., 2015). The marbled crayfish displayed the highest genome wide heterozygosity level, whereas the lowest rates were observed in the morphologically identical *P. fallax*.

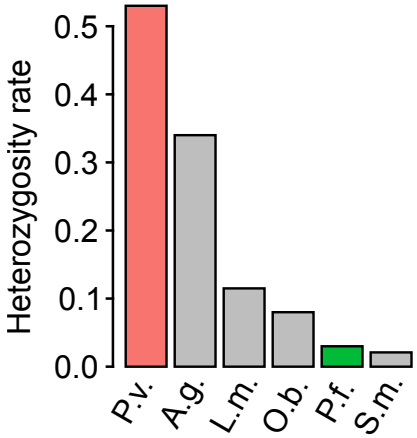

Figure 3.10: Heterozygosity rate in different animals. The estimated heterozygosity rate of *P. virginalis* (P.v., colored in red) compared to other published genomes with likewise inferred levels. Heterozygosity rates are provided as the percentage of heterozygous positions in the genome. Additional organisms include *A. geniculata* (A.g.) (Sanggaard et al., 2014), *L. migratoria* (L.m) (Wang et al., 2014), *O. bimaculoides* (O.b.) (Albertin et al., 2015), and *S. mimosarum* (S.m.) (Sanggaard et al., 2014). A preliminary assembly was generated to approximate heterozygosity levels in *P. fallax* (P.f., colored in green).

3.4.4 Sequence similarity

Variant calling for ten animals was performed to analyze sequence similarity between different *Procambarus* species. Therefore, pairwise comparisons of shared polymorphic loci (to the reference genome) were counted and a maximum distance matrix was calculated. Counts were normalized by the number of total variants for each animal and plotted in an unrooted phylogenetic tree (Figure 3.11). Species relationships can be observed by three distinct clusters, where *P. virginalis* is closer related to *P. fallax* than to *P. alleni*. Moreover, the very short branch lengths for the five marbled crayfish animals (Petshop, Heidelberg, Moosweiher, Illinois, MA1) resulted from almost identical genomes. Substantially more sequence variation was observed in the four *P. fallax* specimens (Male1, Male6, Female1, Female4).

Analysis of Madagascar samples

Five *P. virginalis* samples from Madagascar (MA1, MA2, MA3, MA4, and MA5) were analyzed to illustrate clonality within geographically separated populations. Therefore, single

base substitutions to the reference genome were visualized (Figure 3.11B). Furthermore, variant data from the reference individual (Petshop) and one high coverage *P. fallax* individual (Female4) was used for comparison. Again, sequence similarity in *P. virginalis* animals was observed among all eight genome sequences (comprising 1.5 Mbp of DNA). In the marbled crayfish, some substitutions occur which result from technical artifacts due to low filtering stringency. Considerably more substitutions are observed in *Procambarus fallax*.

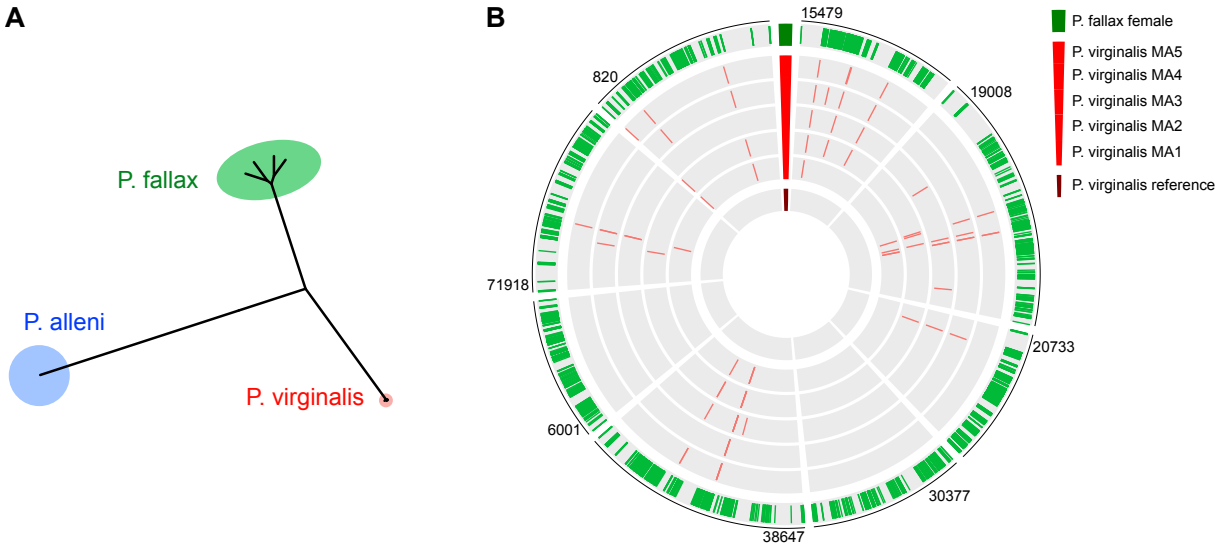

Figure 3.11: Sequence variation in *Procambarus* animals. Sequence variation between different *Procambarus* animals. (A) Phylogenetic tree construction based on pairwise comparison of ten *Procambarus* animals (*P. virginalis* specimens Petshop, Heidelberg, Moosweiher, MA1, Illinois; *P. fallax* specimens Male1, Male6, Female1, Female; *P. alleni* specimen Female1). (B) Ring plot of polymorphisms in 8 randomly chosen marbled crayfish genome sequences. Every ring represents one animal. *P. virginalis* Petshop (dark red) was provided as a reference and *P. fallax* Female4 (green) for comparison. Marbled crayfish populations from Madagascar (colored in red) comprises specimens MA1, MA2, MA3, MA4, and MA5 from five distinct sample sites. Within each ring, a vertical bar represents one single base substitution to the reference genome. Numbers surrounding the outermost ring correspond to the respective scaffold numbers.

3.4.5 Alternative base observations

Alternative base observations provide detailed insights into ploidy of an organism. When analyzing the ratio of alternative to total base observations, the marbled crayfish shows a

major peak at a distribution around 0.33 (Figure 3.12). In other terms, 33% of the mapped reads contain an alternative base at a certain site, also known as heterozygous position. A clearly less emphasized peak can be observed at 0.66, and a negligible amount of position is polymorphic (1.0). More prominent peaks at 0.5 and 1.0 are seen in the diploid *P. fallax* and *P. alleni*. These evidences strongly support a triploid genotype in *P. virginalis*. In particular, a major distribution of one allele being different from the other two alleles suggests an AA'B genotype. As a control, *P. fallax* and *P. alleni* both provide expected evidences for a diploid genotype with 0.5 representing some heterozygous loci and 1.0 being true polymorphisms to the *P. virginalis* reference genome.

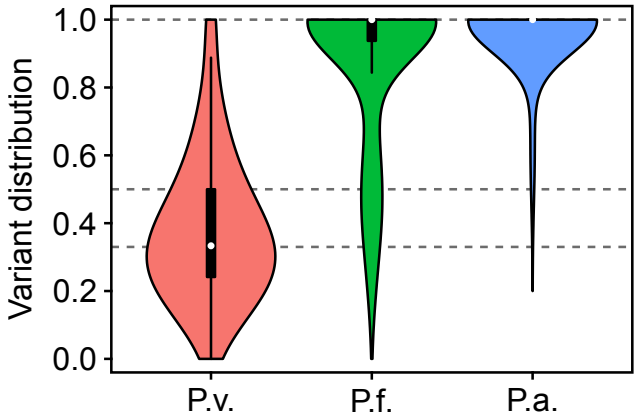

Figure 3.12: Ratio of alternative allele reads. Violin plots of the combined ratios of biallelic alternative loci for three species: *P. virginalis* (red), *P. fallax* (green), and *P. alleni* (blue). Dashed lines are at distributions of 0.33, 0.5, and 1.0. Median values are provided as white dots within boxplots (black bar in each violin). Distributions of 0.33 indicate that 33% of mapped reads contain an alternative single base substitution. An alternative allele frequency of 1.0 represents true polymorphisms to the reference sequence.

3.4.6 Biallelic and triallelic variants

Specific genotypes in a polyploid individuum can elucidate evolutionary processes. Therefore, variants were separated into two groups, biallelic and triallelic variants (Figure 3.13). Variants are considered biallelic when the predicted genotype consists of only two distinct alleles, whereas triallelic variants require three distinct alleles. Valid variations comprise homozygous and heterozygous variants. When comparing profiles of biallelic and triallelic

variants, the first obvious results were found in *P. fallax* and *P. alleni*. There, triallelic variations were entirely absent as they possess only a diploid genotype. Nevertheless, in *P. virginalis* over 99% of heterozygous variants were biallelic rather than triallelic (0.15%). Together with previous results on alternative base distributions, this further suggests an AA'B genotype (due to the ratio of alternate alleles) and not three independently evolved alleles (ABC).

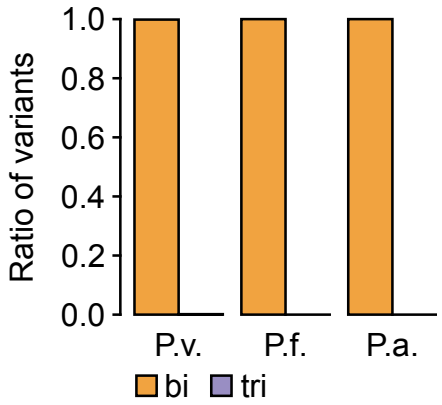

Figure 3.13: Distribution of biallelic and triallelic variants. Average distribution of biallelic and triallelic variants in species *P. virginalis*, *P. fallax*, and *P. alleni*. Triallelic variants (violet) were absent in *P. fallax* and *P. alleni* due to their diploid genotype.

3.4.7 Single base substitutions and mutation signatures

Single base substitutions were analyzed in eight different *P. virginalis* animals (specimens Heidelberg, Moosweiher, Illinois, MA1, MA2, MA3, MA4, and MA5). Prior to estimation, technical artifacts were filtered by remapping reference genome reads (from specimen Petshop). Detailed filtering on polymorphisms is described in the methods section (Section 2.6.4). After filtering, a total of 238 single base substitutions to the reference genome (within all eight animals) were observed. Individual substitution frequencies are provided in the appendix (Appendix Table 6.3).

Subsequently, mutational signatures were investigated to test whether substitutions occur in distinct patterns. A definite peak was observed in substitutions from C to T, followed by G to A and A to G. Less observed were substitutions from A to C and C to G. To further characterize mutations, numerical definitions on substitution patterns were calculated as

previously described (Keith et al., 2016). The 238 mutations were categorized by transitions (130) and transversions (108) resulting in a Ts:Tv (transitions to transversions) ratio of 1.2. Furthermore, the AT:GC ratio was calculated by counting substitutions from C/G→A/T divided by A/T→C/G. In marbled crayfish, an AT:GC ratio of 1.69 was observed.

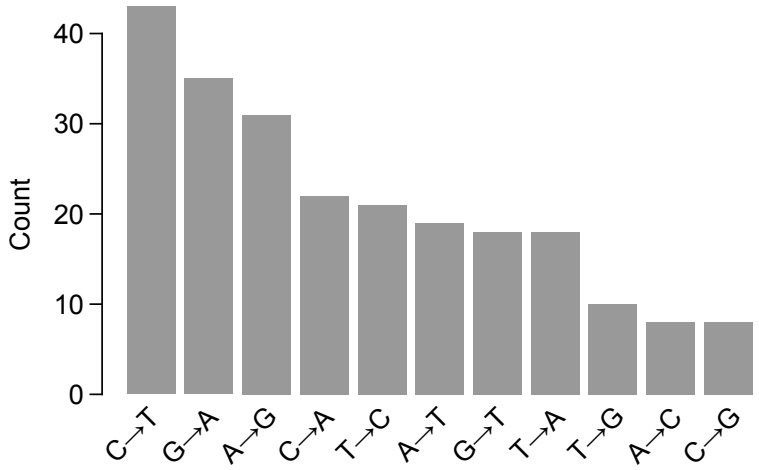

Figure 3.14: Mutation signatures in marbled crayfish. Mutation signatures derived from single base substitutions in eight different *P. virginalis* animals (Heidelberg, Moosweiher, Illinois, MA1, MA2, MA3, MA4, and MA5). Substitutions were combined for all animals.

4 Discussion

With more than 60,000 described species, crustaceans represent a major part within the phylum of arthropods. Specifically, many species are considered ecological and economical keystone species (Crandall and Buhay, 2008). Surprisingly, genome information is still lacking despite their prevalent importance. Among crustaceans, the all-female freshwater crayfish *P. virginalis* shows various unique characteristics such as obligate parthenogenesis and triploidy (Scholtz et al., 2003; Vogt et al., 2004, 2015; Martin et al., 2016). Further favorable traits render the marbled crayfish a promising new model to study a variety of biological principles (Vogt, 2008b). This thesis introduces a first assembly of the marbled crayfish genome and proves clonality in distant populations, hence confirming origination from a single animal.

4.1 Marbled crayfish genome project

Previously, genomic analyses in marbled crayfish was done using extensive laboratory approaches. Thus, the marbled crayfish genome project was initiated in 2013 with the ultimate goal to establish a reference genome as resource for the scientific community.

4.1.1 The Marbled crayfish – a challenging *de novo* genome assembly

The development of eukaryotic genome assembly tools rapidly advanced with projects such as the human genome project and genetic model organisms like *C. elegans*, *A. thaliana*, and *D. melanogaster* (McPherson et al., 2001; Venter et al., 2001; C. elegans Sequencing Consortium, 1998; Arabidopsis Genome Initiative, 2000; Adams et al., 2000). Since then, substantial decrease in sequencing costs and efforts enabled projects for other species.

Projects progressed quickly for organisms such as the honey bee *A. mellifera*, migratory locust *L. migratoria*, bed bug, tick, and spiders (Honeybee Genome Sequencing Consortium, 2006; Wang et al., 2014; Gulia-Nuss et al., 2016; Rosenfeld et al., 2016; Sanggaard et al., 2014). Nevertheless, with most of eukaryotic genomes being diploid many algorithms were developed for diploid genomes with a low heterozygosity (Luo et al., 2012; Zerbino and Birney, 2008; Simpson et al., 2009; Gnerre et al., 2011). High heterozygosity levels required a more sophisticated approach for the large (3.5 Gbp) and triploid marbled crayfish genome.

A genome assembly of competitive quality

The marbled crayfish genome assembly was produced entirely on Illumina sequencing data. While this seems challenging for such a large and complex genome, the strategy of sequencing long jumping distance mate pairs substantially increased genome completeness and sequence representation. Nevertheless, *de novo* sequencing projects often require prior estimations on genome structure, e.g. by low coverage short read sequencing and k-mer estimation. Counting sub strings (k-mers) and plotting depth versus frequency not only allows for a pilot estimation of genome size but also to infer heterozygosity levels, repeat content, and potential sequencing difficulties (Figure 3.1).

The marbled crayfish genome assembly is mainly burdened by high heterozygosity as a result from a triploid genotype. Well-established short read assemblers were not specifically built to perform well on highly heterozygous, polyploid, or repeat rich genomes. Only recently, new approaches were developed to overcome certain limitations caused by heterozygous and repeat rich genomes (Kajitani et al., 2014) or for assembling different haplotypes (Aguiar and Istrail, 2013). Unfortunately, technical requirements for these assemblers were still not as optimized as those of well-established assemblers and hardware requirements often exceeded available resources. In the scope of this project, SOAPdenovo2 was used as it provides compatibility with most systems and a fast run time.

Deduplication of sequences was performed to reduce input complexity. A common problem in the assembly of heterozygous sequences appears upon building de Brujin graph structures. Every alternative allele creates a new path which in turn increases complexity for path traversal. However, the number of paths is certainly limited when reducing the input by deduplication. Although some genomic information was lost, completeness of gene

space was still comparable to other recently published genomes (Figure 3.3B).

In further regard to quality, scaffolding of most sequences succeeded. However, some sequences were just the result from two short contigs spanning long distances. The sequence in between is filled with gap bases. These sequences are considered artifacts as they do not contain actual sequence content and potentially skew assembly statistics. Also, limitations in gap closing algorithms restrict improvements to shorter gaps. Results showed an exponential decrease in extended gap bases for each iteration (Table 3.2). Consequently, an increase in iterations would not substantially improve gap statistics.

Automatic annotation reveals basic gene models and repeat structures

Automatic annotation identified about 21,000 genes. This number is probably underestimated due to fragmentation of the assembly. As the BUSCO benchmark suggests, some genes are entirely missing or fragmented (Figure 3.3B). Gene structures are sometimes scattered in multiple sequences or remain unassembled, and thus they cannot be annotated. Despite these challenges the number of genes is still comparable to other crustaceans (about 30,000 in *D. pulex* and about 28,000 in *P. hawaiensis*) (Colbourne et al., 2011; Kao et al., 2016). In addition, average lengths of genetic features are in the range of both species. This suggests that the quality of automatic annotation is sufficient for comparative analyses. Confirmation of cellulase gene provides an example of usable information in assembly and annotation. Digestion of plant based foods has previously been shown in a variety of freshwater crayfish (Byrne et al., 1999; Crawford et al., 2005). Confirmation of a beta-1,4-glucanase gene verifies gene abundance in marbled crayfish and ultimately demonstrates applicability of the annotation.

The number of *de novo* predicted repeats is also assumed to be slightly underestimated, due to similar reasons as for gene annotation. The genome wide repeat content of 8.8% is lower than reported in other arthropod genomes (9.4% in *D. pulex*, 9.5% in *A. mellifera*, and 14.8-17.9% in bumblebees) (Colbourne et al., 2011; Elsik et al., 2014; Sadd et al., 2015). Fragmentation in genome assemblies often occurs in intergenic or intronic regions. This is due to contamination of repeats, low complexity regions, or other challenging factors for the assembly process. Thus, most automatically annotated repeats are located within or in close proximity of transcript structures.

Interestingly, high abundance of tRNA genes coding for threonine and valine were predicted (Figure 3.7). Total tRNA isotype abundance in eukaryotes, predicted in the same approach, are generally just between 170 and 570 (Goodenbour and Pan, 2006). Whether this increase in tRNA genes is due to genome fragmentation or has a biological meaning remains to be investigated.

It is inevitable that automatic genome annotation should be refined by manual curation. As common practice in *de novo* genome annotation, manual curation potentially improves automatic annotations based on biological data from laboratory work. Therefore, several tools exist permitting researchers or annotation committees to easily review annotations. For this project, a web server for the marbled crayfish genome project was established. The web server provides news about the project, a blast server, and a manual curation tool (http://marmorkrebs.dkfz.de). Ultimately, it provides an extensive central resource enabling researches to improve and work on the annotation of the marbled crayfish genome.

In the last years, rapid advancement in sequencing technologies were made. Especially long read technologies from companies such as PacBio or Oxford Nanopore proved to be useful in *de novo* genome sequencing. These techniques generate reads up to several Kilobases which substantially improves assembly completeness. The application of these methods as well as benefits and drawbacks are further discussed in the outlook (Section 4.2).

In summary, the algorithmic genome annotation is suitable for first comprehensive analyses. Annotation of genetic features is comparable to other recently finished genome projects. Nevertheless, further improvements on assembly structure and annotation are required. The marbled crayfish web server provides a central resource to manually curate annotations.

4.1.2 Clonal evolution and parthenogenesis

Parthenogenesis in arthropods, particularly in crustaceans, was already described (Werren et al., 1995; Lorenzo-Carballa and Cordero-Rivera, 2009; Colbourne et al., 2011). However, parthenogenesis was always considered as an additional form of reproduction, e.g. caused by parasitic infections (Werren et al., 1995). In decapod crustaceans, exclusive reproduction by obligate parthenogenesis is only known in the marbled crayfish. So far, this

was shown by analyses on genetic markers (Scholtz et al., 2003; Vogt, 2011a; Vogt et al., 2015). Here, parthenogenesis in *P. virginalis* is shown on a genome wide scale (Figure 3.11). Phylogenetic analysis on sequence similarity illustrates the relationship of three *Procambarus* species (Figure 3.11A). Conclusively, the marbled crayfish is more closer related to its closes relative *P. fallax* than to *P. alleni*. This supports theories on a direct evolution of the marbled crayfish from *Procambarus fallax* as previously suggested (Vogt et al., 2015; Martin et al., 2016). Pet trade and anthropogenic releases enabled the marbled crayfish to establish populations in different geographic regions (Figure 1.1B). Polymorphism profiles of five animals from Madagascar were analyzed to further elucidate clonal evolution in distantly evolved populations (Figure 3.11B). Analyses clearly show an independent clonal evolution within geographically isolated populations. Furthermore, clonal genomes strongly support an origination from a single origin.

Permanent random mutations in an organism occur during evolution by replication errors or upon DNA damage. The mutation rate for a species provides a measure for mutation accumulation per generation. Marbled crayfish, despite being genetically identical, also undergo mutational events. A total of 238 single base substitutions have been found within eight analyzed animals. The low number of sequence variation further supports and evolution from a single origin. However, inferring a mutation rate for *P. virginalis* was not possible since generation differences between the reference (Petshop) and other animals were not documented. Nevertheless, mutation rate in marbled crayfish is expected to be similar as reported in other arthropod species, e.g. $3.5 \times 10^{-}9$ in *Drosophila melanogaster*, $2.9 \times 10^{-}9$ in *Heliconius melpomoene*, or $4 \times 10^{-}9$ in the cyclic asexual *Daphnia pulex* (Keightley et al., 2009, 2015; Keith et al., 2016).

Mutation signatures are characterized by substitutions from the reference to an alternative nucleotide. Previously, they were used to successfully differentiate different cancer types and origins (Alexandrov et al., 2013). Analysis on signatures in independently evolved *P. virginalis* populations may provide insights into different environmental conditions. Signatures in *P. virginalis* were consistent with the findings in *D. melanogaster* (Keightley et al., 2009). However, pilot analyses did not reveal any population differences. Samples from Madagascar showed similar patterns as observed in other populations, suggesting high adaptive capabilities without the influence of somatic mutations. Hence, phenotypic plas-

ticity is caused by epigenetic regulation rather than by genetic changes. Nevertheless, larger sample sizes from distinct populations should be considered for further insights into mutation signatures caused by environmental stimuli.

Mechanisms of parthenogenesis

Although, the reproductive mode of marbled crayfish was previously studied the molecular mechanisms for apomictic parthenogenesis remain unknown (Lukhaup, 2001; Scholtz et al., 2003; Vogt et al., 2004). The all-female marbled crayfish produce genetically identical offspring by (thelytokous) apomictic parthenogenesis, meaning they do not undergo meiosis on oogenesis (Vogt et al., 2004; Martin et al., 2007; Vogt et al., 2008). Moreover, it was shown that the marbled crayfish evolved from the sexually reproducing (diploid) *Procambarus fallax* (Martin et al., 2010a; Vogt et al., 2015; Martin et al., 2016). The relationship of both species was verified on a genome wide scale using phylogenetics based on sequence similarity (Figure 3.11A).

Pilot analysis on genome sizes between *P. virginalis* and *P. fallax* implies a substantial loss of genomic content in the marbled crayfish (estimations in *P. fallax* performed by F. Gatzmann). Evidences for genomic loss and apomixis suggest aberrations in meiosis relevant genes. Interestingly, analyses on meiosis relevant genes in *P. virginalis* did show irregularities within some structures. However, insights from the transcriptome showed that fragmentation of genes is a limiting factor for precise analyses (Figure 3.3B). Thus, improvements in genome assembly and gene representations are essential for more detailed insights. Furthermore, verification by laboratory experiments should be considered.

Evolution by autopolyploidization

Polyploidy, or more specifically triploidy, is essential for establishing parthenogenesis in marbled crayfish. The ploidy level was already characterized by karyograms and genetic markers (Vogt et al., 2015; Martin et al., 2016). Analysis on heterozygous variant distributions were conducted to verify triploidy on a genome wide level. As expected, the distribution of alternative alleles clearly indicates a ratio of 33% (Figure 3.12). Additionally, it is evident that the marbled crayfish shows an altered distribution pattern when compared to other species of its genus, further consolidating a unique genomic structure.

Upon observing variation patterns, heterozygous positions were rather biallelic than triallelic (Figure 3.13). Triallelic variations require three independently evolved alleles (ABC genotype). By contrast, biallelic variations, the most common type in *P. virginalis*, require only two distantly evolved alleles, of which one is complemented by a highly similar allele (AA'B genotype). In other words, an AA'B genotype arises when the diploid set of alleles from an animal is complemented by an allele from a more distant one. Triploidy and an AA'B genotype strongly favors an evolution by autopolyploidization as previously suggested (Vogt et al., 2015; Martin et al., 2016). Therein, evolution of marbled crayfish was proposed as a macromutation event in captive *P. fallax*. Macromutations could be caused by a heat or cold shock, such that an undifferentiated diploid egg was fertilized by a haploid sperm (Vogt et al., 2015).

Heterozygosity levels in *P. virginalis* are rather high compared to other organisms. Elevated heterozygosity levels were previously reported in *A. geniculata* which also has a very fragmented genome assembly (Sanggaard et al., 2014). Assembly fragmentation could influence heterozygosity levels in marbled crayfish. Nevertheless, high heterozygosity could also be caused by hybridization from two genomically very distant *P. fallax* populations. This further supports an evolution by macromutation. In contrast, very low heterozygosity levels in *P. fallax* were estimated. Here, estimations could be influenced by the very preliminary *P. fallax* genome assembly. Nevertheless, low heterozygosity could also result from inbreeding as they were obtained from a pet trade.

Asexual reproduction and long-term species survival

The benefits and drawbacks of sexual and asexual reproduction have widely been discussed (Lewis, 1987; Crow, 1994; Butlin, 2002). Sexual reproduction is the most prominent form in animals, despite the involved energetic costs and risks (e.g. by infections from sexually transmitted diseases or being exposing during mating). By contrast, asexual reproduction is associated with fewer risks and especially favorable when mating partners are scarce. However, negative aspects of asexual reproduction are a substantial reduction of genetic diversity and slow environmental adaptation. Since only one genotype is passed to the offspring, reproducing by obligate parthenogenesis potentially accumulates deleterious mutations. This negatively affects long-term survival of the species.

Remarkably, some bdelloid rotifers reproduce by obligate apomictic parthenogenesis since 80-100 million years (Mark Welch and Meselson, 2000; Butlin, 2002). Investigation on evolution led to the theory that accumulation of mutations in one locus is independent of the allele in apomictic lineages, which was termed the Meselson effect (Birky, 1996; Mark Welch and Meselson, 2000; Butlin, 2002). Considering the Meselson effect, high heterozygosity and polyploidy could play key factors in long term survival of the marbled crayfish. Evolution of lineages by independent accumulation of mutations (per allele) potentially provides enough genetic diversity to account for the impact of deleterious mutations. Moreover, epigenetic regulation of alleles could play an important role to compensate for any genetic losses (Comai, 2005; Song and Chen, 2015).

Fitness advantage by polyploidy endangers local economy and endemic crayfish on Madagascar

Whole genome duplication in plants is often associated with a significant increase in fitness, survival, and invasive capabilities (Comai, 2005; Otto, 2007; te Beest et al., 2012). Interestingly, some of these characteristics are also observed in the marbled crayfish. Comparisons of biological traits to *P. fallax* show an increase in body size, weight, reproduction, and adaptation to environmental conditions (Jones et al., 2009; Vogt et al., 2015). Moreover, its unique genomic characteristics favors rapid growth of new populations. Studies showed that the marbled crayfish is an invasive species, potentially endangering endemic crayfish in Madagascar, Japan, and other countries (Jones et al., 2009; Kawai and Takahata, 2010; Faulkes et al., 2012). Especially countries with economic dependency in rice production suffer from its rapid spread. Madagascar is a particular example where *P. virginalis* was introduced in 2000 (Jones et al., 2009). There, rice is an important ecological and economical resource. As the climate conditions in some regions are optimal, their fast reproduction allows them to quickly spread across the country. As source for energy they feast on rice plants eventually destroying entire fields. This thesis verifies populations on Madagascar being *Procambarus virginalis*. Interestingly, analyses on mutation events suggest these populations derived from a single origin.

Marbled crayfish favor habitats such as ponds, rivers, swamps, and lakes (Chucholl et al., 2010). Optimal regions and conditions were previously predicted (Feria et al., 2011;

Faulkes et al., 2012). High reproduction rates pose a serious threat for endemic crayfish species, as competition for nutritional sources and habitat space increases. Moreover, the marbled crayfish can act as a host for the crayfish plague, *Aphanomyces astaci* (Unestam, 1972; Steyskall, 2013; Keller et al., 2014). Recently, infected wild populations of marbled crayfish were identified (Keller et al., 2014). Rapid spread and competition for habitats enable the marbled crayfish to uncontrollably transmit the disease. The plague affects several crayfish species eventually leading to death of the animals. Although, the infection is not critical for survival of marbled crayfish, it poses a serious threat for local biodiversity in freshwater habitats (Keller et al., 2014).

4.2 Outlook

In summary, the *de novo* assembled genome of the marbled crayfish is an essential resource for genome wide analyses. Its quality is comparable with recently published arthropod species. Furthermore, it provides a major representative of the genomically very unknown family of crustaceans. Genome assembly is considered an iterative process. Especially, next generation sequencing technologies from PacBio or MinION provide suitable methods to improve assembly structures. Moreover, some questions about genomic characteristics and evolution in the marbled crayfish remain to be investigated.

Iterative approaches to improve *de novo* genome assembly

De novo genome assemblies are rarely considered complete due to technical limitations during the assembly process. Thus, constant improvements in technologies and approaches render algorithmic assembly an iterative process. Examples are the human genome with currently 18 iterations, the *Mus musculus* assembly with 10 iterations, and *Drosophila melanogaster* with 4 iterations. Even smaller genome projects such as the water flea *Daphnia pulex* are constantly being improved (Ye et al., 2017).

The quality of the first marbled crayfish assembly is comparable to other recently published arthropod species (Figure 3.3B). Nevertheless, future iterations should further improve structures such as genes and gap regions (Figure 3.3B and Table 3.2). Assembly refinement can be approached with different methods. Currently, long read sequencing

provides useful information to resolve assembly fragmentation and improve intergenic and intronic regions.

In recent years, advancements in sequencing technologies were made. Two platforms play a major role in *de novo* genome sequencing of larger eukaryotic animals, namely PacBio single molecule real time sequencing (SMRT) and Oxford Nanopore MinION sequencing. Both techniques produce long reads of up to several thousand Kilobases. Therefore, different approaches are pursued.
MinION sequencing (Oxford Nanopore) takes advantage of flow cells containing nano-sized holes, so called nanopores. An ionic current is passed through these nanopores and changes in the current are recorded when a molecule passes through. These changes are then interpreted as base calls, as every nucleotide consists of unique chemical properties. MinION sequencing devices are cost efficient and designed to fit in very small device dimensions. This allows for mobility and potentially finds use in field studies. Lengths of the reads range from 30 kbp to 150 kbp depending on library preparation.
In SMRT sequencing (PacBio) each SMRT cell contains thousands of zero-mode waveguides (ZMW) which are responsible for light detection. Additionally, a DNA polymerase is fixed at the bottom of each ZMW. When DNA fragments are introduced, the DNA polymerase produces a natural DNA strand by incorporating phospholinked nucleotides. When the phosphate is cleaved, a fluorophore is released. Eventually, light emission is measured and translated into base calls. Reads are consistently $>$10 kbp with some even exceeding 60 kbp.

Currently, sequencing accuracy is limited by the technology and chemicals used in long read sequencing. Illumina short read sequencing still provides higher depth and accuracy at much lower costs. To eliminate sequencing errors, long read technologies require high depth sequencing when used for *de novo* genome assembly. Moreover, Illumina is constantly improving its technologies. In 2017, Illumina releases their new NovaSeq system which will replace their current HiSeq protocol. Albeit, producing short reads, the amount of output per run is estimated to be over several thousand Gigabases. Huge amount of data, lower costs, and lower sequencing errors still provide a solid alternative to long read sequencing. Especially, most laboratories already are familiar with Illumina technologies and protocols. Moreover, complementary devices and approaches can still be used. Thus,

new technologies require more time to be adapted and improved. Nevertheless, projects could immensely benefit from a hybrid approach. Combining sequencing data obtained from short and long read technologies could substantially increase assembly quality, especially for complex genome structures (Zimin et al., 2017). Moreover, bioinformatic solutions for hybrid assemblies are already being developed. They combine data from different sequencing platforms to increase accuracy and decrease artifacts (Zimin et al., 2013, 2017; Warren et al., 2015).

Choosing the correct sequencing platform for a new *de novo* genome sequencing project remains a challenging task. Illumina, as it is already established in many laboratories, is still widely used. However, new technologies provide additional options, especially for *de novo* genome sequencing. Hybrid assemblies combine the best of both worlds. Choosing the most suitable sequencing strategy always depends on available resources and genomic characteristics. Nevertheless, it is an essential step when initiating a new genome project.

Biological relevance of the marbled crayfish

The importance of *P. virginalis* as a new model organism and its potential as a threat in freshwater systems was elucidated in previous sections. The outlook further suggests work on genomic characterization and evolution of the marbled crayfish, which was not pursued in the scope of this thesis.

Discussions about new sequencing technologies suggest substantial improvements in assembly completeness. These improvements not only refine genomic structures, but also allow for more detailed analyses on annotation features, allele variation, and species evolution. So far, gene and repeat structures are fairly limited in quality. More complete annotations may provide detailed knowledge about the distribution of transposable elements (TE). Earlier, TE distributions were correlated with obligate parthenogenesis in arthropods (Sullender and Crease, 2001; Eads et al., 2012; Bonandin et al., 2016). For example, transposon insertion in a meiosis relevant gene was described in an obligate parthenogenetic form of *D. pulex* (Eads et al., 2012). Nevertheless, identification on transposable elements requires a more detailed analysis of repeat structures and comprehensive comparisons to the sexually reproducing *P. fallax*. Transposon locations and loads could play a key role in

marbled crayfish to elucidate the mechanisms of obligate apomictic parthenogenesis. Pilot analysis on the distribution of tRNA genes revealed a substantial increase in abundance. As previously discussed, these findings could result from the fragmentation of the assembly. Whether automatic annotation is biased by assembly artifacts or abundance of tRNA genes can be correlated with biological features remains to be investigated. Automatic annotation definitively improves from a more complete assembly.

Evolutionary mechanisms of parthenogenesis in the marbled crayfish remain largely unknown. Speculations were made, but understanding the concepts requires additional work. The assembly of *P. virginalis* enables genome wide analyses in specific genetic regions. Thus, extensive comparative analyses to the morphological identical *Procambarus fallax* should be considered. Understanding the evolution from the diploid *P. fallax* to the triploid and parthenogenetically reproducing marbled crayfish may elucidates genomic mechanisms of obligate parthenogenesis. Pilot analysis on meiosis relevant genes indicated potential mechanisms by structural aberrations. Nevertheless, more insights could be gained by a representative genome assembly of *P. fallax*.

5 Additional contributions

During the period of this doctoral thesis additional contributions to projects in cancer epigenetics were made. Cancer epigenetics describes the field of research dedicated to the analysis of epigenetic mechanisms in cancer formation. Particularly, in eukaryotes methylation of Cytosines (5mC) is an essential mechanism involved in regulating gene expression and chromatin remodeling. In mammalian genomes, DNA methylation is mostly found within CpG dinucleotides, with the Cytosines from both strands being methylated (Bird, 1986). Aberrations in methylation patterns are considered a major hallmark of human cancer.

Whole genome bisulfite sequencing (WGBS) was the first method to analyze genome wide methylation levels on a single base resolution (Frommer et al., 1992). WGBS relies on bisulfite treatment of target DNA converting unmethylated Cytosines to Uracil, which in turn are interpreted as Thymines upon sequencing. Eventually, mapping sequencing results to the genome allows to identify methylated Cytosines within the analyzed samples.

In the study of Abu-Remaileh et al. (2015), chronic inflammation in colitis-induced mouse colons was analyzed for conservation of methylation patterns between intestinal adenomas and colorectal cancer to obtain insights into risk factors of inflammation in tumor formation. Using WGBS, distinct methylation patterns were observed which can be used as markers to correctly classify human colorectal cancer samples. Methylation patterns regulate silencing of active genes involved in gastrointestinal homeostatis. Personal contributions to this project comprised the analysis of expression and co-expression from three identified marker genes (Nr5a2, Ech1, and Foxp2) in wildtype and colon cancer samples. Gene expression levels (normalized counts) were extracted from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). Moreover, methylation differences in single CpG sites within these genes (obtained from Illumina HumanMethylation450 BeadChips) were illustrated.

Besides WGBS, a more cost-efficient technology to detect human DNA methylation is the Illumina HumanMethylation BeadChip. This chip is designed to contain various CpG markers (probes) in carefully selected regions of the human genome. The first two iterations of the chip allowed to interrogate 27 thousand (HumanMethylation27 BeadChip) and 450 thousand (HumanMethylation450 BeadChip) CpG sites, respectively. However, the newest chip (MethylationEPIC BeadChip) contains over 850 thousand sites with additional focus on enhancer regions. Low input requirements and simultaneous analysis of up to 12 samples render HumanMethylation BeadChips a cost-efficient alternative to WGBS.
Sample preparation for HumanMethylation chips also requires bisulfite treatment of the target DNA. Next, treated DNA is amplified (whole genome amplification), fragmented, and denatured before being hybridized onto the BeadChip. Strands are extended by single base incorporation of hapten labeled dideoxynucleotides. Single nucleotide extension is disrupted by unmethylated Cytosines. Immunohistochemical assays and scanning of the chips (red and green channels) results in methylation intensities per site.

During the period of this doctoral thesis a computational pipeline has been developed to automatically process and analyze results from HumanMethylation chips. The pipeline is mainly based on the R Bioconductor package minfi (Aryee et al., 2014). In brief, raw red and green channels from Illumina machines are read, processed, and converted into methylation values. Low quality sites are filtered and normalized for beta value distribution. A linear Bayes regression, as implemented in minfi, is applied to investigate differentially methylated probes. Quality control is provided after each processing step. Results comprise various analyses on differentially methylated probes, beta value distributions, and correlations. The pipeline contributed to three additional projects.

In a study of Geyh et al. (2016) the contribution of mesenchymal stromal cells (MSC) in acute myeloid leukemia (AML) patients was investigated. Personal contributions include the analysis of clearly defined methylation changes between MSC/AML and control samples, which affects pathways involving cell differentiation, proliferation, and skeletal development. The study further found reversible deficits in osteogenic differentiation and insufficient stromal support. Concluding remarks were that MSCs derived from AML samples show molecular and functional alterations and contribute to hematopoietic insufficiency.

Age-related DNA methylation changes have been analyzed using HumanMethyla-

tion450 chips by Bormann et al. (2016). A large scale of human epidermis samples in various age groups were analyzed for global and local methylation patterns. Personal contributions comprise the analysis of differentially methylated regions using the Illumina HumanMethylation pipeline with the result that local age-dependent changes were observed. The study continued to identify discontinuous changes and concludes with an overall reduced epigenetic regulation as consequence of an aging epigenome.

A study of Molina-Pinelo et al. (2016) analyzed methylation patterns within the DLK1-DIO3 cluster of lung cancer patients. Interestingly, methylation changes were observed in lung cancer patients with smoking history, but not in patients without smoking history or COPD patients. Three genes within the DLK1-DIO3 cluster and several non-coding RNA genes were differentially methylated. Personal contributions included the validation of these findings on methylation chip data (HumanMethylation450) extracted from TCGA. In summary, the study suggested the deregulation of the DLK1-DIO3 cluster plays a potential role in the pathogenesis of lung cancer.

5.1 List of publications containing personal contributions

Abu-Remaileh, M., Bender, S., Raddatz, G., Ansari, I., Cohen, D., Gutekunst, J., Musch, T., Linhart, H., Breiling, A., Pikarsky, E., Bergman, Y., and Lyko, F. (2015). Chronic inflammation induces a novel epigenetic program that is conserved in intestinal adenomas and in colorectal cancer. *Cancer Res*, 75(10):2120–30.

Geyh, S., Rodríguez-Paredes, M., Jäger, P., Khandanpour, C., Cadeddu, R.-P., Gutekunst, J., Wilk, C. M., Fenk, R., Zilkens, C., Hermsen, D., Germing, U., Kobbe, G., Lyko, F., Haas, R., and Schroeder, T. (2016). Functional inhibition of mesenchymal stromal cells in acute myeloid leukemia. *Leukemia*, 30(3):683–91.

Bormann, F., Rodríguez-Paredes, M., Hagemann, S., Manchanda, H., Kristof, B., Gutekunst, J., Raddatz, G., Haas, R., Terstegen, L., Wenck, H., Kaderali, L., Winnefeld, M., and Lyko, F. (2016). Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. *Aging Cell*, 15(3):563–71.

Molina-Pinelo, S., Salinas, A., Moreno-Mata, N., Ferrer, I., Suarez, R., Andrés-León, E., Rodríguez-Paredes, M., Gutekunst, J., Jantus-Lewintre, E., Camps, C., Carnero, A., and Paz-Ares, L. (2016). Impact of DLK1-DIO3 imprinted cluster hypomethylation in smoker patients with lung cancer. *Oncotarget*.

Theissinger, K., Falckenhayn, C., Blande, D., Toljamo, A., Gutekunst, J., Makkonen, J., Jussila, J., Lyko, F., Schrimpf, A., Schulz, R., and Kokko, H. (2016). De novo assembly and annotation of the freshwater crayfish Astacus Astacus transcriptome. *Mar Genomics*, 28:7–10.

6 Appendix

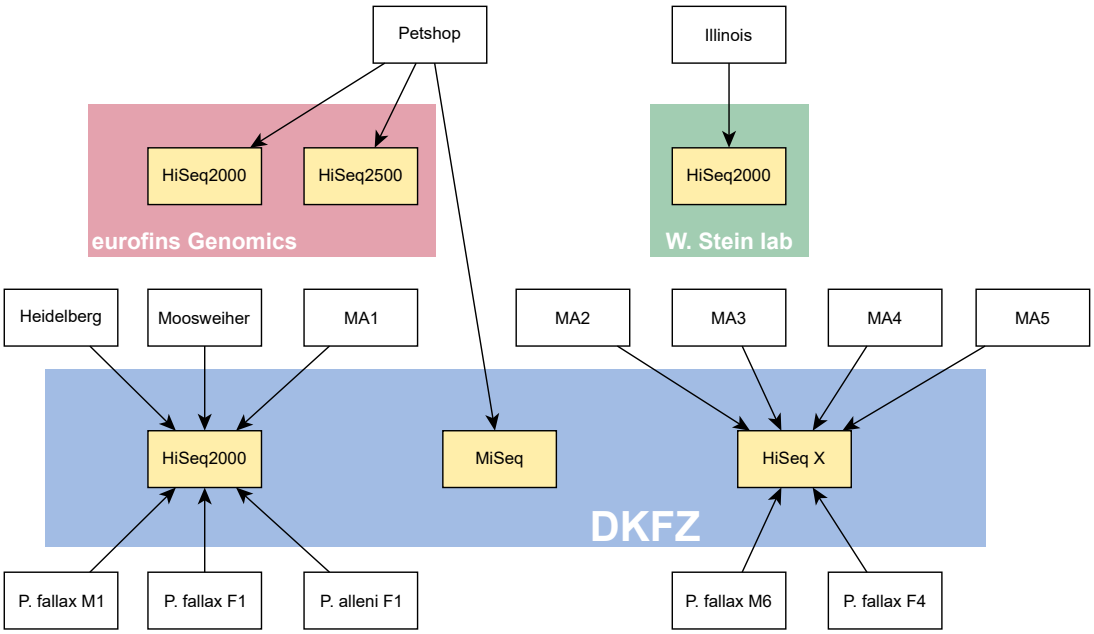

Figure 6.1: Sequencing institutions. Animals (white boxes) used in the scope of this project were sequenced in three independent Institutions. Institutions include Eurofins MWG GmbH (colored in red, from Ebersberg, Germany), the DKFZ Genomics and Proteomics Core Facility (colored in blue, from Heidelberg, Germany), and the W. Stein laboratory (colored in green, from Illinois State University, USA). Institutions provided data from different sequencing technologies (yellow boxes).

Table 6.1: Number of annotation features per source. MAKER algorithms and their contribution (by feature count) to automatic annotation.

source	count
blastn	209,310
blastx	182,785
est2genome	216,601
evm	4,774
maker	230,676
protein2genome	143,922
repeatmasker	481,851
repeatrunner	2,462
snap	151,587
snap_masked	123,889

Table 6.2: Sequencing yield for *Procambarus* animals. Shotgun sequencing yield and read pairs for sequenced *Procambarus* specimens. Read pairs are provided in millions ($\times 10^6$) and raw yield in Mbp (10^6 bp).

ID	Species	Specimen	Read pairs ($\times 10^6$)	Yield (Mbp)
PV1	*P. virginalis*	Petshop	817	245,157
PV2	*P. virginalis*	Heidelberg	669	200,656
PV3	*P. virginalis*	Moosweiher	213	42,937
PV5	*P. virginalis*	Illinois	111	35,581
PV4	*P. virginalis*	MA1 (Ant)	194	39,090
PV6	*P. virginalis*	MA2 (Ala)	374	112,996
PV7	*P. virginalis*	MA3 (Ana)	413	124,643
PV8	*P. virginalis*	MA4 (Ita)	377	113,695
PV9	*P. virginalis*	MA5 (Vak)	361	109,057
PF1	*P. fallax*	Male1	240	48,423
PF2	*P. fallax*	Male6	420	126,845
PF3	*P. fallax*	Female1	206	41,667
PF4	*P. fallax*	Female4	405	122,253
PA1	*P. alleni*	Female1	195	39,314

Table 6.3: Polymorphism frequencies in marbled crayfish samples. Individual polymorphism counts for the eight marbled crayfish specimens.

ID	Species	Specimen	Polymorphism count
PV2	*P. virginalis*	Heidelberg	31
PV3	*P. virginalis*	Moosweiher	12
PV5	*P. virginalis*	Illinois	6
PV4	*P. virginalis*	MA1 (Ant)	23
PV6	*P. virginalis*	MA2 (Ala)	45
PV7	*P. virginalis*	MA3 (Ana)	33
PV8	*P. virginalis*	MA4 (Ita)	49
PV9	*P. virginalis*	MA5 (Vak)	39

Bibliography

Abu-Remaileh, M., Bender, S., Raddatz, G., Ansari, I., Cohen, D., Gutekunst, J., Musch, T., Linhart, H., Breiling, A., Pikarsky, E., Bergman, Y., and Lyko, F. (2015). Chronic inflammation induces a novel epigenetic program that is conserved in intestinal adenomas and in colorectal cancer. *Cancer Research*, 75(10):2120–30.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I.,

Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195.

Aguiar, D. and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29:i352–i360.

Albertin, C. B., Simakov, O., Mitros, T., Wang, Z. Y., Pungor, J. R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C. W., and Rokhsar, D. S. (2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, 524:220–224.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Alwes, F. and Scholtz, G. (2006). Stages and other aspects of the embryology of the parthenogenetic Marmorkrebs (Decapoda, Reptantia, Astacida). *Development genes and evolution*, 216:169–184.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815.

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30:1363–1369.

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nature reviews. Genetics*, 14:333–346.

Bird, A. P. (1986). Cpg-rich islands and the function of dna methylation. *Nature*, 321:209–213.

Birky, C. W. (1996). Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics*, 144:427–437.

Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F., and Apweiler, R. (2002). Applications of InterPro in protein annotation and genome analysis. *Briefings in bioinformatics*, 3:285–295.

Bohman, P., Edsman, L., Martin, P., and Scholtz, G. (2013). The first Marmorkrebs (Decapoda: Astacida: Cambaridae) in Scandinavia. *BioInvasions Records*, 2(3):227–232.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30:2114–2120.

Bonandin, L., Scavariello, C., Mingazzini, V., Luchetti, A., and Mantovani, B. (2016). Obligatory parthenogenesis and TE load: *Bacillus* stick insects and the R2 non-LTR retrotransposon. *Insect science*.

Bormann, F., Rodríguez-Paredes, M., Hagemann, S., Manchanda, H., Kristof, B., Gutekunst, J., Raddatz, G., Haas, R., Terstegen, L., Wenck, H., Kaderali, L., Winnefeld, M., and Lyko, F. (2016). Reduced DNA methylation patterning and transcriptional connectivity define human skin aging. *Aging Cell*, 15(3):563–71.

Braasch, I., Gehrke, A. R., Smith, J. J., Kawasaki, K., Manousaki, T., Pasquier, J., Amores, A., Desvignes, T., Batzel, P., Catchen, J., Berlin, A. M., Campbell, M. S., Barrell, D., Martin, K. J., Mulley, J. F., Ravi, V., Lee, A. P., Nakamura, T., Chalopin, D., Fan, S., Wcisel, D., Cañestro, C., Sydes, J., Beaudry, F. E. G., Sun, Y., Hertel, J., Beam, M. J., Fasold, M., Ishiyama, M., Johnson, J., Kehr, S., Lara, M., Letaw, J. H., Litman, G. W., Litman, R. T., Mikami, M., Ota, T., Saha, N. R., Williams, L., Stadler, P. F., Wang, H., Taylor, J. S., Fontenot, Q., Ferrara, A., Searle, S. M. J., Aken, B., Yandell, M., Schneider, I., Yoder, J. A., Volff, J.-N., Meyer, A., Amemiya, C. T., Venkatesh, B., Holland, P. W. H., Guiguen, Y., Bobe, J., Shubin, N. H., Di Palma, F., Alföldi, J., Lindblad-Toh, K., and Postlethwait, J. H. (2016). The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nature genetics*, 48:427–437.

Bushnell, B. (2015). *BBMap/BBTools*. http://sourceforge.net/projects/bbmap/.

Butlin, R. (2002). Evolution of sex: The costs and benefits of sex: new insights from old asexual lineages. *Nature reviews. Genetics*, 3:311–317.

Byrne, K. A., Lehnert, S. A., Johnson, S. E., and Moore, S. S. (1999). Isolation of a cDNA encoding a putative cellulase in the red claw crayfish *Cherax quadricarinatus*. *Gene*, 239(2):317–324.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282:2012–2018.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17:540–552.

Chucholl, C., Pfeiffer, M., et al. (2010). First evidence for an established Marmorkrebs (Decapoda, Astacida, Cambaridae) population in Southwestern Germany, in syntopic occurrence with Orconectes limosus (Rafinesque, 1817). *Aquatic invasions*, 5(4):405–412.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219.

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Cáceres, C. E., Carmel, L., Casola, C., Choi, J.-H., Detter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., Fröhlich, T., Geiler-Samerotte, K. A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E. V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J. R., Muller, J., Pangilinan, J., Patwardhan, R. P., Pitluck, S., Pritham, E. J., Rechtsteiner, A., Rho, M., Rogozin, I. B., 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, Y. I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J. R., Andrews, J., Crease, T. J., Tang, H., Lucas, S. M., Robertson, H. M., Bork, P., Koonin, E. V., Zdobnov, E. M., Grigoriev, I. V., Lynch, M., and Boore, J. L. (2011). The ecoresponsive genome of *Daphnia pulex*. *Science*, 331:555–561.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature reviews. Genetics*, 6:836–846.

Crandall, K. A. and Buhay, J. E. (2008). Global diversity of crayfish (Astacidae, Cambaridae, and Parastacidae—Decapoda) in freshwater. *Freshwater Animal Diversity Assessment*.

Crawford, A. C., Richardson, N. R., and Mather, P. B. (2005). A comparative study of cellulase and xylanase activity in freshwater crayfish and marine prawns. *Aquaculture research*, 36(6):586–592.

Crow, J. F. (1994). Advantages of sexual reproduction. *Developmental genetics*, 15:205–213.

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

Eads, B. D., Tsuchiya, D., Andrews, J., Lynch, M., and Zolan, M. E. (2012). The spread of a transposon insertion in Rec8 is associated with obligate asexuality in *Daphnia*. *Proceedings of the National Academy of Sciences of the United States of America*, 109:858–863.

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J. D., Foster, L. J., Graur, D., Guigo, R., production teams, H., Hoff, K. J., Holder, M. E., Hudson, M. E., Hunt, G. J., Jiang, H., Joshi, V., Khetani, R. S., Kosarev, P., Kovar, C. L., Ma, J., Maleszka, R., Moritz, R. F. A., Munoz-Torres, M. C., Murphy, T. D., Muzny, D. M., Newsham, I. F., Reese, J. T., Robertson, H. M., Robinson, G. E., Rueppell, O., Solovyev, V., Stanke, M., Stolle, E., Tsuruda, J. M., Vaerenbergh, M. V., Waterhouse, R. M., Weaver, D. B., Whitfield, C. W., Wu, Y., Zdobnov, E. M., Zhang, L., Zhu, D., Gibbs, R. A., and Consortium, H. B. G. S. (2014). Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC genomics*, 15:86.

Falckenhayn, C. (2017). *The Methylome of the Marbled Crayfish Procambarus virginalis*. PhD thesis, Ruperto-Carola University of Heidelberg, Germany.

Faulkes, Z., Feria, T. P., and Muñoz, J. (2012). Do Marmorkrebs, *Procambarus fallax* f. *virginalis*, threaten freshwater Japanese ecosystems? *Aquatic biosystems*, 8:13.

Feria, T. P., Faulkes, Z., et al. (2011). Forecasting the distribution of Marmorkrebs, a parthenogenetic crayfish with high invasive potential, in Madagascar, Europe, and North America. *Aquatic Invasions*, 6(1):55–67.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89:1827–1831.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *bioRXiv*.

Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity*, 100:659–674.

Geyh, S., Rodríguez-Paredes, M., Jäger, P., Khandanpour, C., Cadeddu, R.-P., Gutekunst, J., Wilk, C. M., Fenk, R., Zilkens, C., Hermsen, D., Germing, U., Kobbe, G., Lyko, F., Haas, R., and Schroeder, T. (2016). Functional inhibition of mesenchymal stromal cells in acute myeloid leukemia. *Leukemia*, 30(3):683–91.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, 108:1513–1518.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274:546, 563–546, 567.

Goodenbour, J. M. and Pan, T. (2006). Diversity of tRNA genes in eukaryotes. *Nucleic acids research*, 34(21):6137–6146.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59:307–321.

Gulia-Nuss, M., Nuss, A. B., Meyer, J. M., Sonenshine, D. E., Roe, R. M., Waterhouse, R. M., Sattelle, D. B., de la Fuente, J., Ribeiro, J. M., Megy, K., Thimmapuram, J., Miller, J. R., Walenz, B. P., Koren, S., Hostetler, J. B., Thiagarajan, M., Joardar, V. S., Hannick, L. I., Bidwell, S., Hammond, M. P., Young, S., Zeng, Q., Abrudan, J. L., Almeida, F. C., Ayllón, N., Bhide, K., Bissinger, B. W., Bonzon-Kulichenko, E., Buckingham, S. D., Caffrey, D. R., Caimano, M. J., Croset, V., Driscoll, T., Gilbert, D., Gillespie, J. J., Giraldo-Calderón, G. I., Grabowski, J. M., Jiang, D., Khalil, S. M. S., Kim, D., Kocan, K. M., Koči, J., Kuhn, R. J., Kurtti, T. J., Lees, K., Lang, E. G., Kennedy, R. C., Kwon, H., Perera, R., Qi, Y., Radolf, J. D., Sakamoto, J. M., Sánchez-Gracia, A., Severo, M. S., Silverman, N., Šimo, L., Tojo, M., Tornador, C., Van Zee, J. P., Vázquez, J., Vieira, F. G., Villar, M.,

Wespiser, A. R., Yang, Y., Zhu, J., Arensburger, P., Pietrantonio, P. V., Barker, S. C., Shao, R., Zdobnov, E. M., Hauser, F., Grimmelikhuijzen, C. J. P., Park, Y., Rozas, J., Benton, R., Pedra, J. H. F., Nelson, D. R., Unger, M. F., Tubio, J. M. C., Tu, Z., Robertson, H. M., Shumway, M., Sutton, G., Wortman, J. R., Lawson, D., Wikel, S. K., Nene, V. M., Fraser, C. M., Collins, F. H., Birren, B., Nelson, K. E., Caler, E., and Hill, C. A. (2016). Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature communications*, 7:10507.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838–849.

Holdich, D. M. and Pöckl, M. (2007). Invasive crustaceans in European inland waters. In *Biological invaders in inland waters: Profiles, distribution, and threats*, pages 29–75. Springer.

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

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443:931–949.

Hu, Y., Yan, C., Hsu, C.-H., Chen, Q.-R., Niu, K., Komatsoulis, G. A., and Meerzaman, D. (2014). OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data. *Cancer informatics*, 13:13.

i5K Consortium (2013). The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of heredity*, 104:595–600.

Jones, J. P., Rasamy, J. R., Harvey, A., Toon, A., Oidtmann, B., Randrianarison, M. H., Raminosoa, N., and Ravoahangimalala, O. R. (2009). The perfect invader: a parthenogenic crayfish poses a new threat to Madagascar's freshwater biodiversity. *Biological Invasions*, 11(6):1475–1482.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., and Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, 24:1384–1395.

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Semon, M., Grillo, M., Bruce, H., Kumar, S., Siwanowicz, I., Le, A., Lemire, A., Eisen, M. B., Extavour, C., Browne, W. E., Wolff, C., Averof, M., Patel,

N. H., Sarkies, P., Pavlopoulos, A., and Aboobaker, A. (2016). The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *eLife*, 5.

Kato, M., Hiruta, C., and Tochinai, S. (2016). The Behavior of Chromosomes During Parthenogenetic Oogenesis in Marmorkrebs *Procambarus fallax* f. *virginalis*. *Zoological science*, 33:426–430.

Kawai, T. and Takahata, M. (2010). *Biology of Crayfish*. Hokkaido University Press, Sapporo.

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., Davey, J. W., and Jiggins, C. D. (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular biology and evolution*, 32:239–243.

Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome research*, 19:1195–1201.

Keith, N., Tucker, A. E., Jackson, C. E., Sung, W., Lucas Lledó, J. I., Schrider, D. R., Schaack, S., Dudycha, J. L., Ackerman, M., Younge, A. J., Shaw, J. R., and Lynch, M. (2016). High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. *Genome research*, 26:60–69.

Keller, N., Pfeiffer, M., Roessink, I., Schulz, R., and Schrimpf, A. (2014). First evidence of crayfish plague agent in populations of the marbled crayfish (*Procambarus fallax* forma *virginalis*). *Knowledge and Management of Aquatic Ecosystems*, (414):15.

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simão, F. A., Pozdnyakov, I. A., Ioannidis, P., and Zdobnov, E. M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*, 43(Database issue):D250–6.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9:357–359.

Lewis, W. M. (1987). The cost of sex. *Experientia. Supplementum*, 55:33–57.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079.

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, O. A., Leung, F. C.-C., 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, C. C., Lam, T. T.-Y., 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, M. W., 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, T.-W., Yiu, S.-M., 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, G. K.-S., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, 463:311–317.

Lipták, B., Mrugała, A., Pekárik, L., Mutkovič, A., Grul'a, D., Petrusek, A., and Kouba, A. (2016). Expansion of the marbled crayfish in Slovakia: beginning of an invasion in the Danube catchment? *Journal of Limnology*, 75(2).

Lőkkös, A., Müller, T., Kovács, K., Várkonyi, L., Specziár, A., and Martin, P. (2016). The alien, parthenogenetic marbled crayfish (Decapoda: Cambaridae) is entering Kis-Balaton (Hungary), one of Europe's most important wetland biotopes. *Knowledge and Management of Aquatic Ecosystems*, (417):16.

Lorenzo-Carballa, M. O. and Cordero-Rivera, A. (2009). Thelytokous parthenogenesis in the damselfly *Ischnura hastata* (Odonata, Coenagrionidae): genetic mechanisms and lack of bacterial infection. *Heredity*, 103:377–384.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, 25:955–964.

Lukhaup, C. (2001). *Procambarus* sp., der Marmorkrebs–Ein dankbarer Aquarienbewohner. *Aquaristik aktuell*, pages 7–8.

Lukhaup, C. and Pekny, R. (2003). *Süßwasserkrebse aus aller Welt*. Dähne Verlag.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam,

T.-W., and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18.

Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.-I., Arima, T., Akita, O., Kashiwagi, Y., et al. (2005). Genome sequencing and analysis of *Aspergillus oryzae*. *Nature*, 438(7071):1157–1161.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27:764–770.

Mark Welch, D. and Meselson, M. (2000). Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science*, 288:1211–1215.

Marten, M., Werth, C., and Marten, D. (2004). Der Marmorkrebs (Cambaridae, Decapoda) in Deutschland–ein weiteres Neozoon im Einzugsgebiet des Rheins. *Lauterbornia*, 50:17–23.

Martin, P., Dorn, N. J., Kawai, T., van der Heiden, C., and Scholtz, G. (2010a). The enigmatic Marmorkrebs (marbled crayfish) is the parthenogenetic form of *Procambarus fallax* (Hagen, 1870). *Contributions to Zoology*, 79(3).

Martin, P., Kohlmann, K., and Scholtz, G. (2007). The parthenogenetic Marmorkrebs (marbled crayfish) produces genetically uniform offspring. *Die Naturwissenschaften*, 94:843–846.

Martin, P., Shen, H., Füllner, G., and Scholtz, G. (2010b). The first record of the parthenogenetic Marmorkrebs (Decapoda, Astacida, Cambaridae) in the wild in Saxony (Germany) raises the question of its actual threat to European freshwater ecosystems. *Aquatic invasions*, 5(4):397–403.

Martin, P., Thonagel, S., and Scholtz, G. (2016). The parthenogenetic Marmorkrebs (Malacostraca: Decapoda: Cambaridae) is a triploid organism. *Journal of Zoological Systematics and Evolutionary Research*, 54(1):13–21.

Marzano, F. N., Scalici, M., Chiesa, S., Gherardi, F., Piccinini, A., Gibertini, G., et al. (2009). The first record of the marbled crayfish adds further threats to fresh waters in italy. *Aquatic Invasions*, 4(2):401–404.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner-McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel,

G., Whittaker, A., Holden, J. L., McCann, O. T., Dunham, A., Soderlund, C., Scott, C. E., Bentley, D. R., Schuler, G., Chen, H. C., Jang, W., Green, E. D., Idol, J. R., Maduro, V. V., Montgomery, K. T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J. H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P. J., Catanese, J. J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V. G., Kirsch, I. R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J. F., Hawkins, T., Myers, R. M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N. E., Cox, D. R., Haussler, D., Kent, W. J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X. N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H. S., Sakaki, Y., Shimizu, N., Asakawa, S., Kawasaki, K., Sasaki, T., Shintani, A., Shimizu, A., Shibuya, K., Kudoh, J., Minoshima, S., Ramser, J., Seranski, P., Hoff, C., Poustka, A., Reinhardt, R., Lehrach, H., and Consortium, I. H. G. M. (2001). A physical map of the human genome. *Nature*, 409:934–941.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11:31–46.

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95:315–327.

Molina-Pinelo, S., Salinas, A., Moreno-Mata, N., Ferrer, I., Suarez, R., Andrés-León, E., Rodríguez-Paredes, M., Gutekunst, J., Jantus-Lewintre, E., Camps, C., Carnero, A., and Paz-Ares, L. (2016). Impact of DLK1-DIO3 imprinted cluster hypomethylation in smoker patients with lung cancer. *Oncotarget*.

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2):275–290.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of *Drosophila*. *Science*, 287(5461):2196–2204.

Novitsky, R. A. and Son, M. O. (2016). The first records of Marmorkrebs [*Procambarus fallax* (Hagen, 1870) f. *virginalis*](Crustacea, Decapoda, Cambaridae) in Ukraine. *Ecologica Montenegrina*, 5:44–46.

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, 131:452–462.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Patoka, J., Buřič, M., Kolář, V., Bláha, M., Petrtỳl, M., Franta, P., Tropek, R., Kalous, L., Petrusek, A., and Kouba, A. (2016). Predictions of marbled crayfish establishment in conurbations fulfilled: Evidences from the Czech Republic. *Biologia*, 71(12):1380–1385.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reese, M. G. and Guigó, R. (2006). EGASP: Introduction. *Genome biology*, 7 Suppl 1:S1.1–S1.3.

Richman, N. I., Böhm, M., Adams, S. B., Alvarez, F., Bergey, E. A., Bunn, J. J., Burnham, Q., Cordeiro, J., Coughran, J., Crandall, K. A., et al. (2015). Multiple drivers of decline in the global status of freshwater crayfish (Decapoda: Astacidea). *Phil. Trans. R. Soc. B*, 370(1662):20140060.

Rieger, V. and Harzsch, S. (2008). Embryonic development of the histaminergic system in the ventral nerve cord of the Marbled Crayfish (Marmorkrebs). *Tissue & cell*, 40:113–126.

Rosenfeld, J. A., Reeves, D., Brugler, M. R., Narechania, A., Simon, S., Durrett, R., Foox, J., Shianna, K., Schatz, M. C., Gandara, J., Afshinnekoo, E., Lam, E. T., Hastie, A. R., Chan, S., Cao, H., Saghbini, M., Kentsis, A., Planet, P. J., Kholodovych, V., Tessler, M., Baker, R., DeSalle, R., Sorkin, L. N., Kolokotronis, S.-O., Siddall, M. E., Amato, G., and Mason, C. E. (2016). Genome assembly and geospatial phylogenomics of the bed bug *Cimex lectularius*. *Nature communications*, 7:10164.

Sadd, B. M., Barribeau, S. M., Bloch, G., de Graaf, D. C., Dearden, P., Elsik, C. G., Gadau, J., Grimmelikhuijzen, C. J. P., Hasselmann, M., Lozier, J. D., Robertson, H. M., Smagghe, G., Stolle, E., Van Vaerenbergh, M., Waterhouse, R. M., Bornberg-Bauer, E., Klasberg, S., Bennett, A. K., Câmara, F., Guigó, R., Hoff, K., Mariotti, M., Munoz-Torres, M., Murphy, T., Santesmasses, D., Amdam, G. V., Beckers, M., Beye, M., Biewer, M., Bitondi, M. M. G., Blaxter, M. L., Bourke, A. F. G., Brown, M. J. F., Buechel, S. D., Cameron, R., Cappelle, K., Carolan, J. C., Christiaens, O., Ciborowski, K. L., Clarke, D. F., Colgan, T. J., Collins, D. H., Cridge, A. G., Dalmay, T., Dreier, S., du Plessis, L., Duncan, E., Erler, S., Evans, J., Falcon, T., Flores, K., Freitas, F. C. P., Fuchikawa, T., Gempe, T., Hartfelder, K., Hauser, F., Helbing, S., Humann, F. C., Irvine, F., Jermiin, L. S., Johnson, C. E., Johnson, R. M., Jones, A. K., Kadowaki, T., Kidner, J. H., Koch, V., Köhler, A., Kraus, F. B., Lattorff, H. M. G., Leask, M., Lockett, G. A., Mallon, E. B., Antonio, D. S. M., Marxer, M., Meeus, I.,

Moritz, R. F. A., Nair, A., Näpflin, K., Nissen, I., Niu, J., Nunes, F. M. F., Oakeshott, J. G., Osborne, A., Otte, M., Pinheiro, D. G., Rossié, N., Rueppell, O., Santos, C. G., Schmid-Hempel, R., Schmitt, B. D., Schulte, C., Simões, Z. L. P., Soares, M. P. M., Swevers, L., Winnebeck, E. C., Wolschin, F., Yu, N., Zdobnov, E. M., Aqrawi, P. K., Blankenburg, K. P., Coyle, M., Francisco, L., Hernandez, A. G., Holder, M., Hudson, M. E., Jackson, L., Jayaseelan, J., Joshi, V., Kovar, C., Lee, S. L., Mata, R., Mathew, T., Newsham, I. F., Ngo, R., Okwuonu, G., Pham, C., Pu, L.-L., Saada, N., Santibanez, J., Simmons, D., Thornton, R., Venkat, A., Walden, K. K. O., Wu, Y.-Q., Debyser, G., Devreese, B., Asher, C., Blommaert, J., Chipman, A. D., Chittka, L., Fouks, B., Liu, J., O'Neill, M. P., Sumner, S., Puiu, D., Qu, J., Salzberg, S. L., Scherer, S. E., Muzny, D. M., Richards, S., Robinson, G. E., Gibbs, R. A., Schmid-Hempel, P., and Worley, K. C. (2015). The genomes of two key bumblebee species with primitive eusocial organization. *Genome biology*, 16:76.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467.

Sanggaard, K. W., Bechsgaard, J. S., Fang, X., Duan, J., Dyrlund, T. F., Gupta, V., Jiang, X., Cheng, L., Fan, D., Feng, Y., Han, L., Huang, Z., Wu, Z., Liao, L., Settepani, V., Thøgersen, I. B., Vanthournout, B., Wang, T., Zhu, Y., Funch, P., Enghild, J. J., Schauser, L., Andersen, S. U., Villesen, P., Schierup, M. H., Bilde, T., and Wang, J. (2014). Spider genomes provide insight into composition and evolution of venom and silk. *Nature communications*, 5:3765.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27:863–864.

Scholtz, G., Braband, A., Tolley, L., Reimann, A., Mittmann, B., Lukhaup, C., Steuerwald, F., and Vogt, G. (2003). Ecology: Parthenogenesis in an outsider crayfish. *Nature*, 421:806.

Schurko, A. M., Logsdon, J. M., and Eads, B. D. (2009). Meiosis genes in *Daphnia pulex* and the role of parthenogenesis in genome evolution. *BMC evolutionary biology*, 9(1):78.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., Boore, J. L., Grigoriev, I. V., Lindberg, D. R., Seaver, E. C., Weisblat, D. A., Putnam, N. H., and Rokhsar, D. S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, 493:526–531.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–2.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22:549–556.

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

Song, Q. and Chen, Z. J. (2015). Epigenetic and developmental regulation in plant polyploids. *Current opinion in plant biology*, 24:101–109.

Steyskall, C. (2013). Optimisation of a Real Time PCR detection assay for *Aphanomyces astaci* and its application in host-pathogen interaction studies of Astacus astacus, *Procambarus fallax* forma *virginalis* and *Gammarus* spp. Master's thesis, University of Graz, Austria.

Sullender, B. W. and Crease, T. J. (2001). The behavior of a *Daphnia pulex* transposable element in cyclically and obligately parthenogenetic populations. *Journal of Molecular Evolution*, 53(1):63–69.

te Beest, M., Le Roux, J. J., Richardson, D. M., Brysting, A. K., Suda, J., Kubesová, M., and Pysek, P. (2012). The more the better? The role of polyploidy in facilitating plant invasions. *Annals of botany*, 109:19–45.

Theissinger, K., Falckenhayn, C., Blande, D., Toljamo, A., Gutekunst, J., Makkonen, J., Jussila, J., Lyko, F., Schrimpf, A., Schulz, R., and Kokko, H. (2016). De Novo assembly and annotation of the freshwater crayfish *Astacus astacus* transcriptome. *Marine Genomics*, 28:7–10.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22:4673–4680.

Unestam, T. (1972). On the host range and origin of the crayfish plague fungus. *Rep Inst Freshw Res Drottningholm*, 52:192–198.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew,

R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., 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, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., 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, M. L., 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, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., 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, Y. H., 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., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Vilpoux, K., Sandeman, R., and Harzsch, S. (2006). Early embryonic development of the central nervous system in the Australian crayfish and the Marbled crayfish (Marmorkrebs). *Development genes and evolution*, 216:209–223.

Vogt, G. (2007). Exposure of the eggs to 17alpha-methyl testosterone reduced hatching success and growth and elicited teratogenic effects in postembryonic life stages of crayfish. *Aquatic toxicology*, 85:291–296.

Vogt, G. (2008a). How to minimize formation and growth of tumours: potential benefits of decapod crustaceans for cancer research. *International journal of cancer*, 123:2727–2734.

Vogt, G. (2008b). The marbled crayfish: a new model organism for research on development, epigenetics and evolutionary biology. *Journal of Zoology*, 276(1):1–13.

Vogt, G. (2009). Research on aging and longevity in the parthenogenetic marbled crayfish, with special emphasis on stochastic developmental variation, allocation of metabolic resources, regeneration, and social stress. *Handbook on longevity: genetics, diet and disease. Nova Science Publishers, Hauppauge*, pages 353–372.

Vogt, G. (2010). Suitability of the clonal marbled crayfish for biogerontological research: a review and perspective, with remarks on some further crustaceans. *Biogerontology*, 11:643–669.

Vogt, G. (2011a). Marmorkrebs: natural crayfish clone as emerging model for various biological disciplines. *Journal of biosciences*, 36(2):377–382.

Vogt, G. (2011b). Marmorkrebs: natural crayfish clone as emerging model for various biological disciplines. *J Biosci*, 36(2):377–82.

Vogt, G. (2012). Hidden treasures in stem cells of indeterminately growing bilaterian invertebrates. *Stem Cell Reviews and Reports*, 8(2):305–317.

Vogt, G. (2015). Stochastic developmental variation, an epigenetic source of phenotypic diversity with far-reaching biological consequences. *J Biosci*, 40(1):159–204.

Vogt, G. (2016). Structural specialties, curiosities, and record-breaking features of crustacean reproduction. *J Morphol*, 277(11):1399–1422.

Vogt, G., Falckenhayn, C., Schrimpf, A., Schmid, K., Hanna, K., Panteleit, J., Helm, M., Schulz, R., and Lyko, F. (2015). The marbled crayfish as a paradigm for saltational

speciation by autopolyploidy and parthenogenesis in animals. *Biology Open*, 4(11):1583–94.

Vogt, G., Huber, M., Thiemann, M., van den Boogaart, G., Schmitz, O. J., and Schubart, C. D. (2008). Production of different phenotypes from the same genotype in the same environment by developmental variation. *The Journal of experimental biology*, 211:510–523.

Vogt, G., Tolley, L., and Scholtz, G. (2004). Life stages and reproductive components of the marmorkrebs (marbled crayfish), the first parthenogenetic decapod crustacean. *Journal of morphology*, 261:286–311.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B., and Kang, L. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nature communications*, 5:2957.

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., and Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, 4:35. Original DateCompleted: 20150805.

Werren, J. H., Zhang, W., and Guo, L. R. (1995). Evolution and phylogeny of Wolbachia: reproductive parasites of arthropods. *Proceedings of the Royal Society of London B: Biological Sciences*, 261(1360):55–63.

Xue, W., Li, J.-T., Zhu, Y.-P., Hou, G.-Y., Kong, X.-F., Kuang, Y.-Y., and Sun, X.-W. (2013). L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics*, 14:604.

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

Ye, Z., Xu, S., Spitze, K., Asselman, J., Jiang, X., Ackerman, M. S., Lopez, J., Harker, B., Raborn, R. T., Thomas, W. K., Ramsdell, J., Pfrender, M. E., and Lynch, M. (2017). A New Reference Genome Assembly for the Microcrustacean *Daphnia pulex*. *G3*.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., Seppey, M., Loetscher, A., and Kriventseva, E. V. (2017). OrthoDB v9.1: cataloging

evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*, 45(D1):D744–D749.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18:821–829.

Zhang, G., Rahbek, C., Graves, G. R., Lei, F., Jarvis, E. D., and Gilbert, M. T. P. (2015). Genomics: Bird sequencing project takes off. *Nature*, 522:34.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29:2669–2677.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., and Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome research*, 27:787–792.

Acknowledgements

My sincere thanks go to Prof. Dr. Frank Lyko for the opportunity to pursue this project. He not only provided me this unique possibility, but more importantly supported, encouraged, and inspired me. I greatly enjoyed the talks about science and, of course, our shared passion to American Football.

During the time of my thesis I was very lucky to meet many wonderful persons in the Lyko lab (A130). A heartful thanks to my former office colleagues: Achim, Cassandra, Fanny, and Ranja. You made the long hours of work very pleasant. More so, special thanks to Cassandra and Kathi for helping me so much with my project. Particularly, I also want to thank Matthias. With him I not only gained a very good friend but he also helped me in the process of writing this thesis. Of course, thanks be due to all the people in the lab for the regenerative coffee breaks and nice chats. Muchas gracias to Manolo for always supporting and pushing me. He provided me so many opportunities for additional contributions and publications. And thanks to the Epigenetics running group for keeping me fit and healthy.

I want to thank Prof. Dr. Benedikt Bors for being my second reviewer and TAC member. He and Dr. Oleg Simakov contributed valuable input and comments on this project. Their expertise and experiences provided inspirations for further analyses. Thanks to Wolfgang Stein for collaborating and providing sequencing results from his lab.

Special thanks to Helena for always supporting and helping me wherever she could. She was always there for me and kept my spirits up when I needed her.

And of course, I am very grateful my awesome family. They made everything possible for me and were always supporting and caring. Thanks to my dad and my brothers who were more than often an inspiration. Especially, I want to thank my mom who worked, fought, and cared for me all my life. Without you, I would never have come this far.

Thank you.

,