Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the

Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M. Sc. Olena Maiakovska

born in Tashkent, Uzbekistan

Oral examination: ______

Origination, Monoclonality and Evolution of the Marbled Crayfish Genome *Procambarus virginalis*

Referees

Prof. Dr. Frank Lyko

Prof. Dr. Benedikt Brors

“The mind is not a vessel to be filled, but a fire to be kindled.”
— Plutarch

I dedicate this PhD thesis to my first teacher, my mother.

Abstract

The parthenogenetic marbled crayfish (*Procambarus virginalis*) sparked interest within the scientific community due to its unique features. Its polyploid and monoclonal genome, high environmental adaptability and phenotypic diversity made the marbled crayfish a suitable laboratory model for genomics, epigenetics and ecology research.

The previously established marbled crayfish genome sequence of 3.5 Gbp represents a highly fragmented draft assembly. Initial comparative genomic analyses resulted in confirmation of *P. virginalis* genome origination from the sexually reproducing freshwater crayfish *P. fallax*. However, in-depth genomic analysis and interspecies genome comparisons require further refinement of the fragmented genome reference of the marbled crayfish.

In this PhD thesis, the first refinement of the marbled crayfish genome has been performed with application of the PacBio Single Molecule Real Time (SMRT) sequencing technology. The new and improved genome assembly of the marbled crayfish resulted in 3.7 Gbp of sequence length and an N50 of 144kb. The refined genome assembly enabled searching parental haplotypes and understanding species origination. The absence of evidence for loss of heterozygosity in the various monoclonal marbled crayfish generations suggests the lack of recombination process during oogenesis. Thus, marbled crayfish suggest to be apomictic parthenogens which are characterized by generating identical copies of the maternal genotype. Moreover, despite of the limited genome variability, monoclonal marbled crayfish genomes consisted of population-specific genetic polymorphisms within the global population. Comparative genomic analysis between geographically distant populations resulted in the identification of population-specific mutational signatures. The calculation of genomic variability of marbled crayfish from the growing population in Lake Reilingen allowed to estimate population dynamics. Thus, the population in Lake Reilingen demonstrates a rapid growth, following the density-independent exponential model.

This PhD thesis provides fundamental insights into marbled crayfish research, particularly via making use of an improved genome assembly for comparative genomic analyses, epigenetic studies, and for research on the evolution and genomic adaptation to asexuality.

Kurzzusammenfassung

Aufgrund siner einzigartigen Eigenschaften weckte der parthenogenetische Marmorkrebs (*Procambarus virginalis*) das Interesse innerhalb der Wissenschaft. Sein polyploides und monoklonales Genom, starke Anpassungsfähigkeit sowie phänotypische Veränderungen machen den Marmorkrebs zu einem geeigneten Modellorganismus in genetischer, epigenetischer und ökologischer Forschung.

Bisher wurde der Marmorkrebs durch ein vorläufiges, 3.5 Gbp großes und stark fragmentiertes Referenzgenom beschrieben. Erste genomische Vergleichsanalysen konnten bestätigen, dass sich *P. virginalis* vom, durch sexuelle Fortpflanzung reproduzierenden, Flusskrebs *P. fallax* abstammt. Jedoch braucht es für tiefergehende Analysen eine verbesserte Version des fragmentierten Marmorkrebs Referenzgenoms.

In dieser Doktorarbeit wurde das bisherige Referenzgenom des Marmorkrebs mit Hilfe des PacBio Single Molecule Real Time (SMRT) Sequenzierverfahren schrittweise verbessert. Die neu assemblierte Genomsequenz erreicht eine Gesamtlänge von 3.7 Gbp mit einem N50 Wert von 144 kbp. Weiterhin lässt diese neue Genomsequenz es zu nach den Haplotypen der Elternspezies zu suchen um dadurch die Entstehung einer solchen Spezies zu verstehen. Der Mangel an Beweisen für den Verlust der Heterozygosität in den verschiedenen monoklonalen Marmorkrebsen deutet auf Oogenese ohne Rekombination hin. Somit charakterisiert sich der Marmorkrebs als Organismus der sich durch apomiktische Parthenogenese vermehrt und identische Kopien nur des maternalen Genotyps generiert. Trotz der limitierten genetischen Veränderungen des Genoms weißt der monoklonale Organismus populationsspezifische genetische Polymorphismen in der weltweiten Verbreitung auf. Genomische Vergleiche zwischen geografisch separierten Populationen brachten eindeutige und spezifische Mutationssignaturen hervor. Mittels der Auswertung über genetischen Variabilität des Marmorkrebs innerhalb einer wachsenden Population im Reilinger See konnte die Populationsdynamik erfasst werden. Zum Beispiel zeigte die Population im Reilinger See einen rapiden Wachstum welcher einem Populationsdichte-unabhägigen exponentiellen Modell gleicht.

Diese Doktorarbeit gibt Einblicke in die fundamentale Wissenschaft rund um den Marmorkrebs, inbesondere benutzt es das in dieser Arbeit verbesserte Referenzgenom für Vergleichsanalysen im Bereich der Genetik, Epigenetik, Evolution und der genomische Anpassung zur Asexualität.

List of Abbreviations

A	Adenine
ASD	Allele sequence divergence
bp	Base pair
C	Cytosine
CCS	Circular consensus sequence
CLR	Continuous long read
cm	Centimetre
CNV	Copy number variation
CPU	Central processing unit
DNA	Deoxyribonucleic acid
dNTP	Deoxyribo-nucleoside triphosphate
dsDNA	Double stranded DNA
G	Guanine
Gbp	Gigabase pair
FGS	First generation sequencing
HGP	Human Genome Project
HPC	High performance computing
HTS	High-throughput sequencing
Hi-C	Chromosome conformation capture
LINE	Long interspersed nuclear element
LJD	Long jumping distance
LOH	Loss of heterozygosity
kbp	Kilo base pair
Mbp	Megabase pair
MP	mate-pair
µg	Microgram
N	Ambiguous gap nucleotide
NGS	Next generation sequencing
NJ	Neighbor-Joining
nt	Nucleotide

OLC	Overlap-Layout-Consensus
ONT	Oxford Nanopore Technology
PacBio	Pacific Biosciences
PCA	Principal components analysis
PE	Paired-end
RNA-seq	RNA-sequencing
SG	Shotgun (reads)
SINE	Short interspersed nuclear element
SMRT	Single molecule real time (sequencing)
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SVs	Structural variations
T	Thymine
TE	Transposable element
WGS	Whole-genome sequencing
ZMW	Zero-mode waveguide

Contents

1 Introduction

The genome of each organism consists of information embedded in the complete set of deoxyribonucleic acid (DNA), which is needed to build and maintain the organism. The full length of this DNA sequence can range from several thousand base pairs (bp) (e.g. bacterial genome) to tens of billions of bp (e.g. complex genomes of plants and animals). The genome sequence of each organism is unique and differs between species, as well as between individuals within a single species. The progress in sequencing technologies has enabled the growth of numerous biological fields including evolutionary genomics, biology of speciation, ecology and molecular biology. The availability of genomic DNA sequences has become possible through the advance in sequencing technologies and algorithms developed for the efficient assembly of sequenced reads.

1.1 DNA sequencing technologies and genome assembly

The major finding in biological science in the last century was the description of the DNA double-helix by James Watson and Francis Crick in 1953 (Watson and Crick, 1953). The composition of DNA structure includes four different deoxyribonucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T), which are linked together by a sugar-phosphate backbone. Two backbones are the strands, which are positioned in the opposite directions (5'-to-3' and 3'-to-5') and form a spiral structure. The hydrogen bonds between nucleotides keep the two strands together and form connections according to complementarity. Wherein the nucleotide A is paired with T, and the nucleotide C paired with G. The bonded nucleotides are named as base pairs (bp), which are also used as a unit of DNA length (Chargaff et al., 1951).

1.1.1 First and second generation of DNA sequencing technologies

The first leading technology of DNA sequencing was the Sanger method (or First Generation Sequencing technology, FGS) which provides low-throughput sequences for a high cost (Sanger et al., 1977b). The first genome, sequenced by FGS, was the small genome of PhiX174 bacteriophage (Sanger et al., 1977a). It took many years to sequence another genome of bigger size, such as the 1.8Mb genome of bacterium, *Haemophilus influenzae* (Fleischmann et al., 1995). The biggest achievement was the completion of the Human Genome Project (HGP) in 2001, when the initial sequencing and analysis of the human genome was performed by utilizing the FGS technology and by pioneering assembly algorithms (Venter et al., 2001).

The second-generation sequencing technologies (or Next Generation Sequencing, NGS) appeared in 2000 and became widely available after 2004 (Barba et al., 2014). It included HiSeq from Illumina, 454 Life sciences from Roche, Solexa and SOLiD, which generate short reads that range between 50 and 400 bp in length. Moreover, their main feature is the ability to massively parallelize sequencing reactions. The high-throughput read production and substantial cut in cost of sequencing resulted in the growth of a number of genome projects. However, the short read length was a major limiting factor, which was solved by the application of Paired-End (PE) and Mate-Pair (MP) sequencing. PE and MP are strategies for reading the ends of bigger DNA fragments with different lengths of insert sequence. The insert sequence length of PE and MP can range from 200 bp to 600 bp, and from 2,000 bp to 40,000 bp, respectively. This allows to link two distant read pairs and thus, to increase the contiguity of genome assembly. However, NGS-based genome assemblies remain highly fragmented with long unresolved bases or gap bases. Apart from the application in genome assembly, NGS reads are found application in genotypic determination (Elshire et al., 2011), epigenetic variation (Tsompana and Buck, 2014) or in the detection of small structural variations, such as Single Nucleotide Polymorphisms (SNPs) (Deschamps and Campbell, 2010).

1.1.2 Third generation of DNA sequencing technologies

The recently emerged sequencing technologies are named Third Generation Sequencing (TGS) technologies, and allow the generation of exclusively long reads. TGS technologies include Single Molecule Real Time (SMRT) sequencing from Pacific Biosciences of California, Inc. (PacBio) (Eid et al., 2009) and nanopore sequencing from Oxford Nanopore (Schadt et al., 2010). Both technologies were released in 2011 and 2014 and since then have become advantageous for numerous applications in genome biology.

The SMRT sequencing platform utilizes a sequencing-by-synthesis approach, when DNA polymerase incorporates the fluorescently labelled nucleotides into native DNA strands in real-time fashion. The DNA polymerase is anchored to the bottom of the sequencing unit called zero-mode waveguide (ZMW). ZMW provides the smallest volume for light detection during the synthesis reaction. On 3000 of ZMWs were allocated on a prototype chip called SMRT cell, where the replication process was parallelized in ZMWs. The recorded "movie" of light pulses from each ZMW can be interpreted into a sequence of bases, which represents a continuous long read (CLR) (Eid et al., 2009). The process of sequencing lasts in dependance of the polymerase lifetime and usually reaches the read length of 40 Kb. However, the sequencing error rate is high and reach 15%, which is usual case in generation of TGS reads (Lee et al., 2016). These highly erroneous long reads can be improved by the generation of multiple subreads from CLR via specific modification of DNA fragments. The hairpin adapters are ligated at both ends of the double stranded DNA (dsDNA) fragment and allow the formation of a circular template (Figure 3), which has to be sequenced by DNA polymerase in multiple passes. The subreads with an approximate rate of 15% used for the generation of consensus highly accurate reads, known as "Circular Consensus Sequences" (CCS). The long and accurate reads enabled the identification of small and large sequence variations in the genome, as well as their successful application in *de novo* genome assembly. The recent optimisation of CCS generation resulted in 99.8% accuracy for long high-fidelity reads (HiFi), but with a reduces average read length (13.5 kb) (Wenger et al., 2019). The PacBio company successfully developed the single-molecule real-time sequencing and increased the throughput and the cost-effectivity. Thus, their first Sequel I platform, which has the SMRT cell with 1 M of ZMWs was recently replaced by Sequel II platform, which delivers about 8X more reads with 8 million ZMWs per SMRT Cell.

The Oxford Nanopore sequencing is distinct by principle from SMRT sequencing. It measures the ionic current fluctuations when the single-stranded nucleic acids pass through the biological nanopores in an insulated membrane. Different nucleotides of the DNA strand pass through the pore and block the flow of ions, changing the current and creating a signature of signals. This signature is recorded and translated into specific base calls. The length of sequenced reads depends on the actual length of DNA fragments prepared in the library (Schadt et al., 2010). Similar to SMRT sequencing, nanopore sequencing includes the protocol of sequence refinement via ligation of a hairpin adapter, which allows the sense and antisense stands to be sequenced consecutively. The resulting data includes the generated higher quality consensus sequences ("2D reads"). The initial error rate was reported as 35% (Laver et al., 2015), but later it was decreased with the development of the new version of base callers and nanopore chemistry.

The major application of TGS as a provider of long read sequences is their devotion towards resolving ambiguous assembly gaps caused by repetitive DNA sequences. Together with substantial improvements in chemistry and software development, TGS technologies overcame the early limitations in accuracy and throughput (Wenger et al., 2019; Amarasinghe et al., 2020).

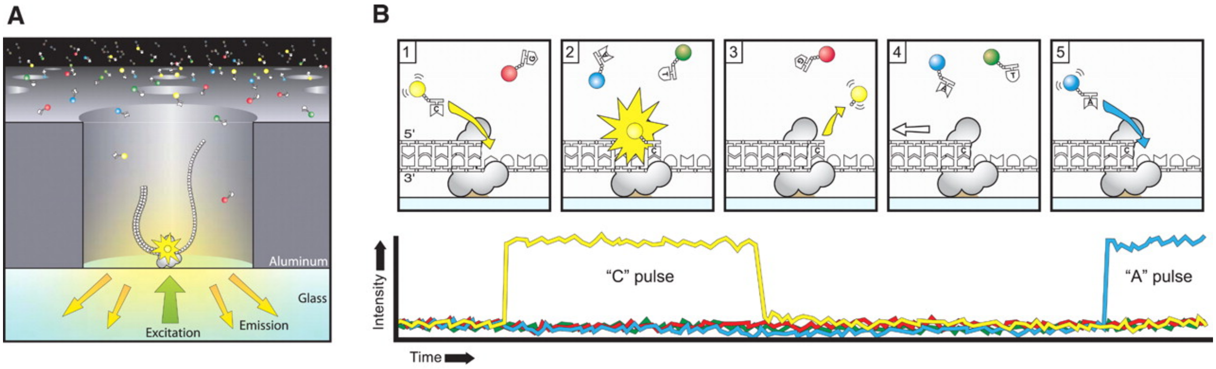

Figure 1: **Principle of single-molecule, real-time DNA sequencing.** A: The zero-mode waveguide (ZMW) well with immobilized polymerase at the bottom. The laser light illumination and emission detection from individual phospholinked nucleotide substrates occur from below of ZMW. B: The cycle of phospholinked Deoxynucleoside Triphosphate (dNTP) incorporation and formation of DNA strand. Four fluorescent-labeled nucleotides generate the distinct emission spectrums while incorporation. (1) Formation of cognate complex with the template in the polymerase active site, (2) emission activation of corresponding spectrum, (3) cleavage of dye-linker-pyrophosphate product with corresponding fall of the fluorescence pulse, (4) the polymerase translocates forward, and product of cleavage diffuses out of the ZMW, (5) the next phospholinked dNTP enters the cycle (Eid et al., 2009).

1.1.3 Genome assembly

A genome assembly is the reconstruction of a genome sequence, starting from DNA fragments and resulting in long chromosome-scale scaffolds. This assembly undergoes iterat-

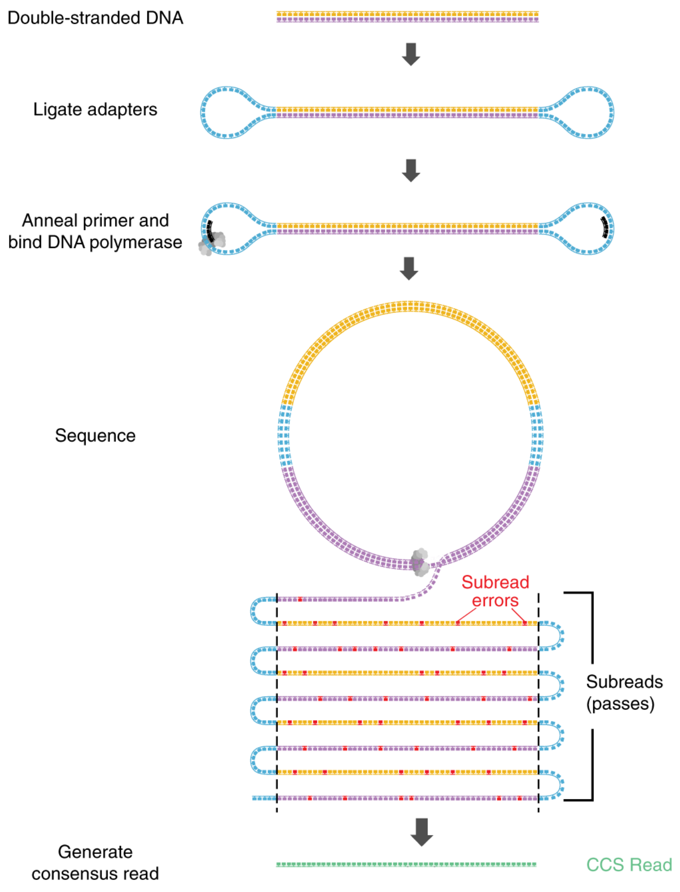

Figure 2: **Generation of circular consensus sequence (CCS).** The adapter sequences (colored in blue) are ligated to the double-stranded DNA (colored in purple and orange) and form the circular DNA sequence. The DNA polymerase via multiple passes through the circular DNA generates the long sequence, which consists of the subreads. These noisy individual subreads can be realigned for building the accurate CCS (colored in green). The image is adapted from Wenger et al., 2019.

ive processes of finding the overlaps from short-fragment sequences, building and joining the contiguous portions of the target genome (contings) into longer scaffolds. The ultimate task in the assembly is the ordering and orientation of scaffolds to chromosome-level.

De novo genome assembly is a fundamental problem in bioinformatics. The goal of *de novo* assembly is to assemble the DNA sequence of an unknown genome from the sequence information of fragments, when neither the draft of reference genome nor the genome of related species are available. The *de novo* human genome assembly was a colossal and expensive work spanning more than a decade (International Human Genome Sequencing Consortium, 2004). Since then, with the development of NGS technologies, such as 454 (Margulies et al., 2005), Illumina (Bentley et al., 2008) sequencing, TGS technologies and substantial improvement of assembly algorithms, *de novo* genome assembly remains the an active topic of genome biology. The wide use of NGS and TGS technologies resulted in numerous short-read based, long-read based, or hybrid genome assemblies (Wheeler et al., 2008; Li et al., 2010; Dalloul et al., 2010; Tan et al., 2018; Xia et al., 2019; Wu et al., 2020) and in the completion of several human resequencing projects (Wang et al., 2008; Schuster et al., 2010; Ju et al., 2011).

Short reads provided by NGS reach the length of hundreds of bases and have impressively low error rate (under 1%), are computationally demanding. For instance, the first and most popular assembly algorithm which was used initially for FGS, Overlap-Layout-Consensus (OLC) was not applicable anymore due to difficulties in finding the overlap between short reads and the impossibility to scale to enormous amount of reads (Pop, 2009). Thus, numerous short-read assemblers have been developed and are based on string data structure, like de Brujin graphs (Chaisson et al., 2004) and String graphs (Myers, 2005) (covered in details in the following paragraphs). These approaches demonstrated a greater time-efficiency and high fidelity in genome reconstruction. However, the big genome size and its structure restrict the successful application of NGS, given that it generally does not provide a complete assembly of a particular genome. The major challenges of NGS application on assembly of genomes are the presence of repetitive sequences and the scalability problem. The presence of repetitive elements in the genome, which appear multiple times in different positions of a genome sequence, makes it difficult to distinguish among several repeated regions and map them to their respective genomic locations. In case, when the repeat region is longer than the read provided by NGS, the identification of genome sequence becomes a challenging task. The scalability problem appears in the case of an enormous amount of read datasets, which translate to the generation of several billion vertices in the de Bruijn graph. This task is computationally demanding requiring high capacities of RAM and is also time-costly. Providing accurate long reads, TGS technologies have become a useful alternative for genome assembly. They successfully deal with large repeats and complex elements in the genome through by covering with a single TGS read the entire length of chromosomes in bacterial genome, complex long repeat elements in eukaryotes, and thus, avoiding the problem of fragment joining during the assembly process.

Depending on the length of sequence fragments the different algorithms are applied for assembly. This is a computationally challenging problem and thus, upgrading sequencing technologies and *de novo* genome assemblers continue to be an active research topic.

Algorithms for genome assembly

The alignment of obtained reads against each other in order to find if there is any overlap between them is not a trivial task. The overlapping regions are the most important for the assembly algorithm. As was mentioned previously, three major approaches, such as Overlap Layout Consensus (OLC), de Bruijn graphs and String graphs are the most commonly used algorithms for genome assembly. The detailed discussion of all three approaches can be found in the manuscripts, which are briefly summarised below (Myers Jr, 2016; Simpson and Pop, 2015; Chaisson et al., 2015b).

OLC relies on the construction of a pairwise alignment and consists of three major steps. The first step is the identification of overlaps between the reads. The second includes the construction of a graph, which is based on a layout of all the reads and their overlaps information. The last step is the construction of a consensus sequence from the graph. All reads are pooled together for finding the overlaps. The algorithm constructs an overlap graph, where reads are considered as nodes and assigns an edge between two nodes only when the overlap reaches the cutoff threshold length. With an increased number of reads due to higher coverage, the number of nodes and edges increases and therefore the OLC method works better for a smaller set of relatively longer reads. It was implemented in Celera assembler, used for the first human genome assembly with FGS data (Denisov et al., 2008; Venter et al., 2001).

The de Bruijn graph method has been adopted by many assemblers and consists of several steps. The first starts with the graph construction based on the overlaps between the decomposed reads of defined length (named as k-mers, k > 0, usually ranged between 31 and 200). The second step includes error correction, when the generated path in the de Bruijn graph undergoes pruning for artifacts caused by sequencing errors and inconsistencies. The last step is the contig generation, when the paths have to be enumerated to build the sequence (the contigs or scaffolds). In the graph contraction step the k-mers correspond to the nodes and the edges are defined between the vertices of any two consecutive k-mers in a read (k-mers should overlap by k-1 nucleotides). de Bruijn graph counts the frequency of k-mer occurrence and thus the coverage is not a restricting factor in effective algorithm work, unlike OLC. CPU and memory usage is substantially decreased in requirements for the assemblers working on de Bruijn graph and thus, the NGS reads of high coverage do not constrain their efficient application for large genome assembly. Many assemblers are based on this approach and there are tools which apply it in combination with other assembly approaches. Several tools utilize the de Bruijn graph in their algorithms for the efficient genome assembly, such as EULER (Pevzner et al., 2001) and Velvet (Zerbino and Birney, 2008) assemblers.

The string graph is the fundamental method of long read based assemblers. The string graph is obtained from an overlap graph by removing redundancy. For example, if two reads A and C overlap, but at the same time both overlap with other read B, such as A–>B and B–>C so then redundancy is applied and results in graph based on A–>C connection. The short reads enclosed in the long reads are removed in the transitive reduction. This approach is memory efficient and similar to the operating principle of OLC. The assemblers, such as FALCON (Chin et al., 2013) or NECAT (Chen et al., 2020) utilize the string graphs for PacBio and Nanopore reads correspondingly.

Despite the advantages of current approaches, the assembly of large and complex genomes still requires a memory-intensive computational performance and a long time to complete the assembly.

Challenges of genome assembly

Despite the progress which has been made in last decades in sequencing technologies and the substantial improvement of assembly algorithms, there are factors commonly affecting the quality of genome assembly:

- **Repetitive sequences.** Repetitive sequences in the genome were the major challenge during the NGS application for the assembly of large genomes. But the existing repeats of different length, such as microsatellites, macrosatellites, centromeric repeats, transposable elements, segmental duplications and other are still represent the major barrier in obtaining the correct and contiguous assembly of repeat-rich genome (Chaisson et al., 2015b).

- **Sequence coverage.** The genome assembly requires an even read coverage over the entire genome length. Although due to technical limitations the sequence coverage can lead to appearance of gaps in the assembly. TGS technology produces the reads which cover evenly the genome regions despite the GC content, unlike NGS reads, which are sensitive to GC content (Lee et al., 2016).

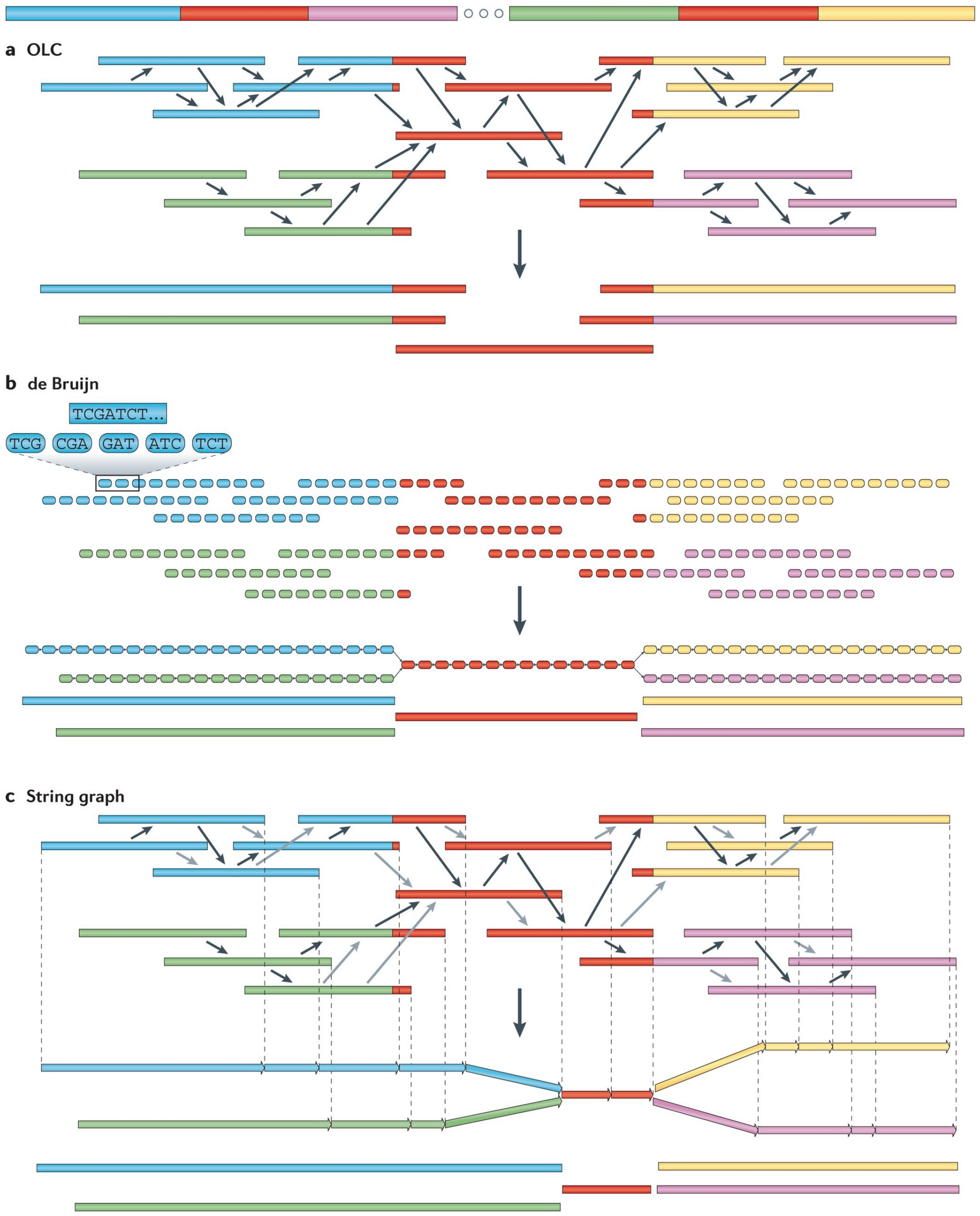

Figure 3: **Three major approaches used in *de novo* genome assembly.** Overlap Layout Consensus (OLC), de Bruijn and String graphs. OLC performes the pairwise alignment between the reads and merges them into consensus sequence. de Bruijn approach based on finding the overlaps between k-mers of decomposed reads. The k-mers corresponding to repeat sequence have to undergo the additional graph operations for the assembly into consensus sequence. String graph removes all redundant alignments (dashed lines) for the efficient generation of consensus sequence. The repeat sequences colored in red. The alternative consensus sequences represented in parallel orientation and differently colored (Chaisson et al., 2015b).

- **Ploidy.** Most of the genome assemblies consist of collapsed haplotypes, representing a 'mosaic' sequence of both (or more) haplotypes. The higher genome ploidy makes the generation of assembly graphs more complicated by adding the extraneous paths (Chaisson et al., 2015b). A variety computational algorithms for genome phasing are developed and applied to resolve ploidy problem in the genome assembly (Zhang et al., 2020).

- **Sequencing errors.** Illumina reads consist of 1-2% of errors, which affect the overlap search in the assembly algorithm and lead to generation of erroneous or extraneous paths (Minoche et al., 2011; Simpson and Pop, 2015). TGS has a higher error rate (15%), but successfully utilized after the intensive correction step. The complexity in the genome assembly algorithms increase with the number of errors despite of read length. Errors occur more often in the GC- or AT-rich regions. Thus, the important step in the genome assembly workflow is the error correction (Au et al., 2012; Liao et al., 2019).

Depending on the length of sequence fragments the different algorithms are applied for assembly. This is a computationally challenging problem and thus, upgrading sequencing technologies and *de novo* genome assemblers continue to be an active research topic.

1.1.4 Genome projects

In era of sequencing, when highly accurate and fast sequencing technologies became more efficient and affordable many genome projects were initiated. Thus, the first genome drafts and the complete assemblies for hundreds of thousands species have become available. The fast and well-established workflow for the assembly enables initiatives providing the pan-genome reference genomes. The examples include The Global Ant Genomics Alliance (Boomsma et al., 2017), The 100,000 genomes project (England, 2016), Earth BioGenome Project (Lewin et al., 2018), The i5K Initiative (i5K Consortium, 2013) and many others. These projects aim to provide a comprehensive dataset for genome diversity within genera and help to find the common genomic features, which represent the biology and evolutionary trajectories of species. Moreover, the availability of genomes of commercially important species brings the input in population genetics, ecology and socio-economic studies.

The first genome projects were initiated mainly for the assembly of organisms with genomes of small size, such as different prokaryotes and viruses. Only in last decades the number of genome projects aimed to assemble the larger genomes of eukaryotes has been increased (Figure 4). Moreover, the recently initiated genome projects mostly include the numbers of genomes of different individuals of one species or representative individuals of certain taxonomic group. Thus, 1000 of human genomes are included in 1000 Genomes Project (Consortium et al., 2010) or G10K initiative for vertebrates genomes assembly (Koepfli et al., 2015) aims to assemble 10,000 genomes of vertebrate species. These large initiatives are important for the analysis of genomic variations in population studies and understanding of evolution and speciation within the taxonomic ranks.

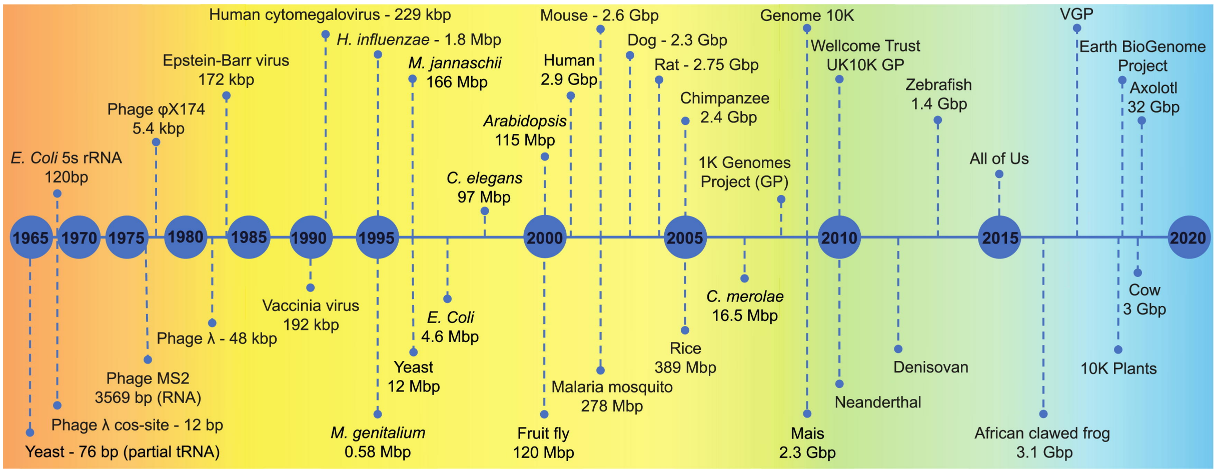

Figure 4: **The major genome projects.** The genome assembly projects started with introduction of early sequencing methods, which were replaced by Sanger-based shotgun sequencing (corresponding period colored in orange-yellow). After 2009, the majority of genome assemblies were NGS-based (green background). TGS technologies have become advantageous in genome assembly application since 2016 (colored in blue) (Giani et al., 2020).

The first genome assembly of marbled crayfish was initiated in 2017 at the division of Epigenetics at German Cancer Research center (Gutekunst et al., 2018). Recently, this animal was introduces as a model organism for epigenetics research due to its high degree of phenotypic plasticity (Carneiro and Lyko, 2020). The genome assembly was in high demand since it represents the starting point of genetic studies. More detailed description on marbled crayfish species and first marbled crayfish genome project given in the next subsections of Introduction.

1.2 Marbled crayfish as a model organism

The introduction of marbled crayfish as a model organism started in the last five years, when its unique biological features become unraveled and its scientific unvalued input has grew.

This animal has become well known very recently, in 1995 the first record of species was found in a German pet shop. It is a popular aquarium pet due to its aesthetic appearance, easy to culture and high fertility (Figure 5). Later in 2003, the first discovery of the parthenogenic nature of marbled crayfish was made, when purely female gonads were established in a morpho-physiological study reporting the first obligate parthenogenetic decapod (Scholtz et al., 2003). Moreover, the mode of reproduction and high adaptive capabilities of marbled crayfish allow it to reproduce up to seven times, with an average 400 number of offspring per clutch (Figure 6)(Seitz et al., 2005). These features bring a high scientific potential for marbled crayfish as a model organism.

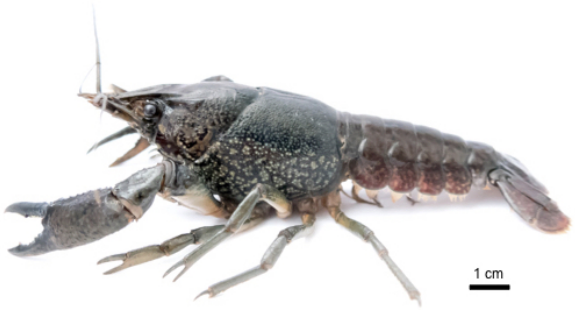

Figure 5: ***Procambarus virginalis* holotype.** Lateral view of specimen of 19 cm (Lyko, 2017)

Since mid-1990s the species has been known only within the German aquarium trade but in recent 5 years it represents not only an exceptionally aquarium populations, but formed invasive wild populations in many European countries and invaded a large territory of Madagascar. The origin of worldwide wild populations assumed the anthropogenic releases and distribution via the pet trade. The wide geographical range of newly established stable populations gives an evidence about high tolerance to diverse ecological niches and certain advantages of marbled crayfish, which potentially can lead to the extrusion of native species. Thus, the confirmed reports with stable marbled crayfish populations include Madagascar, Germany, Slovakia, Czech Republic, Romania, Hungary, Ukraine, Estonia and Malta (Gutekunst et al., 2018; Jones et al., 2009; Pârvulescu et al., 2017; Deidun et al., 2018; Chucholl, 2015; Lipták et al., 2016; Novitsky and Son, 2016; Patoka et al., 2016; Ercoli et al., 2019). In Madagascar, marbled crayfish population has rapidly spread in the last decade, within territory of 100,000 km^2 and nowadays have a commercial impact in the country (Andriantsoa et al., 2019). In the laboratory condition marbled crayfish reproduce all year round (Vogt, 2015) with a generation time approx. 6 months. Since marbled crayfish became a popular laboratory study object, its parthenogenic nature was one of the focus of the study. First, high genetic identity within generations observed in marbled crayfish indicates the presence of apomictic mode of parthenogenic reproduction (Gutekunst et al., 2018). Microsatellite markers covered heterozygous loci in the analysis of various generations revealed identical sequences in marbled crayfish bringing the conclusion that genetic information was not recombined. Secondly, the absence of meiosis during oocyte maturation was assumed in histological studies of nuclei (Vogt et al., 2004) and additionally, the karyotype analysis revealed the triploidy of marbled crayfish genome (Martin et al., 2016). Lately, the histological study with the application of immunohistochemistry brought the first evidence of the presence of meiosis-like incomplete division. The destabilizing chromosome behavior during parthenogenic oogenesis was described and referred to the intrinsic ones in other polyploid pathogenic species (Kato et al., 2016).

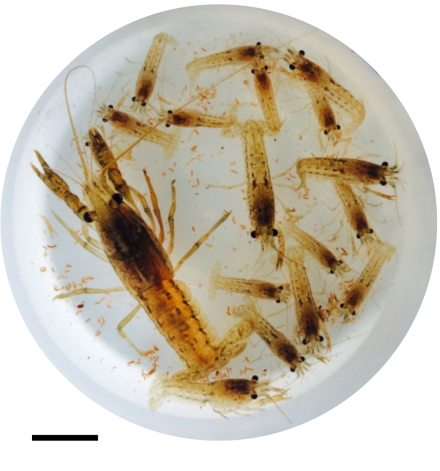

Figure 6: **Marbled crayfish in a single clutch.** Marbled crayfish siblings from the same clutch of eggs displaying different size and coloration patterns. (Gutekunst et al., 2018)

1.2.1 Properties of the marbled crayfish genome

The karyological analysis revealed a triploid genotype of marbled crayfish, consisting of 276 chromosomes, which correspond to triple set of haploid genome of closely related species *P. fallax*. In the following genomic studies of marbled crayfish the size of haploid genome was estimated and resulted in 3.5 Gbp. Moreover, the first draft *de novo* genome assembly of the marbled crayfish genome was released recently (Gutekunst et al., 2018). The assembly was performed based on NGS data and resulted in total sequence length of 3.3Gb, weighted mean sequence length (N50) of 39.4 kb. This first reference genome initiated the comprehensive genomic studies, when the high heterozygosity of genome (0.53%), together with triploid biallelic genome structure (AA'B) were detected (Gutekunst et al., 2018). The genome annotation resulted in identification of more then 21,000 genes, among which about one-fifth part were unique transcripts, which never found n annotation of other related taxonomic groups (Gutekunst et al., 2018).

Further comparative analysis based on whole-genome sequencing of marbled crayfish individuals from the various aquarium lineages and from wild populations in Madagascar firmly supported the genome monoclonality. The marbled crayfish meta-populations are originated form a single, genetically homogeneous clone. The identified genetic variants are suggested to occurred due to natural mutagenesis, but the their number is limited by the extremely young evolutionary age of the species (Gutekunst et al., 2018). After the release of *de novo* draft assembly of the marbled crayfish genome, the other comprehensive genomic studies, transcriptome and methylome analyses were enabled.

Being the unique example of parthenogenic all-female species, marbled crayfish represent an important model for the study of parthenogenesis and genome evolution. The clonal genome of marbled crayfish makes this crustacean particularly suitable for research in epigenetics. The first genome-wide methylome was published and a comprehensive analysis of gene body methylation was reported recently (Falckenhayn, 2017; Gatzmann et al., 2018), which are important for introducing marbled crayfish as a model organism in the epigenetic field.

1.2.2 Origination

Analysis of the morphology, physiology and genome-wide sequencing comparisons revealed a particularly close relationship of marbled crayfish to slough crayfish *Procambarus fallax* (Hagen 1870) (Vogt et al., 2015). Florida and southern Georgia form the natural habitat of slough crayfish, but a record of marbled crayfish was never found at these localities (Scholtz et al., 2003; Martin et al., 2010). *P. fallax* is sexually reproducing species, which was considered to have a parthenogenic form, the marbled crayfish and was named as *Procambarus fallax* forma *virginalis* (Martin et al., 2010). Until recently the marbled crayfish has been considered as a species on its own and named *Procambarus virginalis* (Lyko, 2017). The separation into independent species was enabled after the establishment of reproductive isolation and differences in mitochondrial genomes between marbled crayfish and *P. fallax* (Vogt et al., 2015).

The genomic studies have suggested the model of marbled crayfish genome origination, where the gamete undergoes autopolyploidization and after fertilization it results in formation of triploid organism (Gutekunst et al., 2018). However, the analysis of *P. fallax* haplotypes and their search within triploid genome has not been preformed. Thus, the genetic, phylogeographic, and mechanistic origin of the species have still remained unresolved.

1.2.3 Parthenogenesis

Parthenogenetic reproduction characterised by development of embryos from unfertilized eggs. It has different modes with corresponding distinct outcomes (Mittwoch, 1978). Thus, automixis leads to female gametes fusion after the reductional division with the production of genetically distinctive offspring due to the presence of genetic recombination. Another mode of parthenogenic reproduction is apomixis, which characterized by bypass of meiotic recombination. The result of it leads to generation of clonal progeny, when the offspring are the exact copy of the maternal genotype.

A change in the mode of reproduction from sexual to parthenogenic is expected to have numerous consequences in evolution, ecology and biology of species. The origination of parthenogenesis was a major focus of evolutionary biology due to its importance in evolution of species and its diverse cause. Very frequent reason of parthenogenesis origination has a genetic basis: hybridisation between distantly related populations or different species, mutation is sex-specific genes or changes caused by infection with endosymbionts (Stouthamer et al., 1993; Normark, 2003). The lack of sexual reproduction and active recombination leads to limited phenotypic diversity and environmental adaptation. Interestingly, that parthenogenesis positively correlates with polyploidy, when both also are associated with significant increase in fitness and survival. The explanation for it was given by introducing a term “general-purpose genotype” for parthenogens with the broader ecological niches, successfully surviving in a diverse and fluctuating environment and “frozen niche variation” hypothesis, when parthenogen has the narrower niches and a single genotype inherited from genetically heterogeneous sexual group (Lynch, 1984; Vrijenhoek, 1979). The representative examples are the asexual species from independent generas successfully surviving in a diverse and fluctuating environment (Lynch, 1984; van der Kooi et al., 2017; Goudie et al., 2012). Arising from hybridization and polyploidity, pathenogens increase the heterozygosity and gene redundancy, which shield their genomes from the deleterious effect of mutations (Comai, 2005). Due to these genomic features, a big number of parthenogens evolve to have broadly tolerant genotypes, demonstrating a higher niche diversity than their sexual relatives (van der Kooi et al., 2017). Similarly to that, parthenogenic triploid marbled crayfish demonstrate the adaptability to the wide environmental conditions and in comparison with its sexual parental species *P. fallax*, it shows an increase in body size, weight and reproduction (Vogt et al., 2015; Jones et al., 2009; Martin et al., 2016). It is still unknown if the marbled crayfish could be considered as an example of geographical parthenogenesis described by Vandel (Vandel, 1940) and be a result of a direct consequence of selection for high heterozygosity.

1.3 Aims of this PhD thesis

The marbled crayfish is a unique model organism which has been recently introduced in genomic and epigenetic research. Its parthenogenic nature and polyploid monoclonal genome has been established. Moreover, the high adaptability and phenotypic plasticity of this animal raised a great interest in mechanisms of (epi-)genomic regulations and evolution of species.

The marbled crayfish reference genome draft assembly has been recently published. However, it exhibit a highly fractured sequence and represents the major obstacle in the marbled crayfish genome-related studies.

This doctoral thesis aimed to refine the marbled crayfish reference genome assembly by utilizing the advantageous SMRT sequencing provided by PacBio. Long reads sequencing offers the promising alternatives to resolve genome assembly difficulties, mostly in repetitive regions where the assembly of short-read sequencing data often fails. However, the highly-accurate Illumina reads together with transcriptomic data (RNA-seq) can be introduced to the assembly workflow for error correction of long reads and improvement of scaffolding. The scientific relevance of refined version of marbled crayfish genome assembly includes its application in research aimed to understand the genome origination, evolution and monoclonality. In the framework of genome origination study, this thesis focused on investigation of parental haplotypes of the marbled crayfish among the the closely related sexually reproducing species *P. fallax*. The genome-wide identification of genomic variants within independent clonal marbled crayfish populations intended to determine the population or lineage-specific patterns of mutational activity. The calculation of genomic variability of population can aid in the estimation of marbled crayfish clonal population dynamics.

2 Materials and Methods

This chapter dedicated to description of technical infrastructure, sample preparation, generation of sequencing data, employed quality control procedures, *de novo* genome assembly workflow and details of downstream bioinformatic analyses.

2.1 Computing system

2.1.1 Hardware

Computations for genome assembly were performed on a high-performance cluster running the Slurm Workload Manager using up to 56 CPUs and 450 GB memory. Mapping, trimming and other sequencing data processing were performed on DKFZ's high-performance computing cluster (HPC) with 352 cores on 21 nodes, comprising 2 TB RAM. Downstream analyses were performed on a local desktop computer with eight CPUs in four cores and 24 GB of RAM.

2.1.2 Software packages

Software packages and versions are mentioned in the Table 2. The detailed description of usage commans will be mentioned in the following subsections on this chapter. Various versions were used according to different platforms required for each stage of bioinformatic analysis.

2.2 Sample acquisition, preparation and sequencing

2.2.1 Origin of samples

Animal samples used in this study have various origination and described in the Tables 3 - 4.

- *De novo* genome assembly: three independent *P. virginalis* animals used for the library preparation come from the lineage in our division (Vogt et al., 2015).

- Various *P. virginalis* populations: The specimens were collected by collaboration partners from the natural habitat in various European countries and Madagascar (Table 3). The whole-genome sequencing data of four other specimens from independent populations in Madagascar were obtained from publicly available dataset (Gutekunst et al., 2018).

- *P. fallax* specimens: were provided by collaboration partners Christopher E. Skelton (C.E.S.) and Nathan J. Dorn (N.J.D.) and by several authors of the recent manuscripts (Manteuffel-Ross et al., 2018; Levy et al., 2017). The samples were collected from 23 sites covered southern Georgia and Florida (Table 4).

Table 2: **The basic software packages used in this research.** Various versions were executed in different computational environments.

Software	Version	Source
FastQC	0.11.3	Andrews et al. (2010)
Trimmomatic	0.32	Bolger et al. (2014)
picard	1.137	Tools (2015)
SAMtools	1.3, 1.8	Li et al. (2009)
bowtie2	2.2.6	Langmead and Salzberg (2012)
Freebayes	1.0.2	Garrison and Marth (2012)
R	3.2.3, 3.6.1	Team et al. (2013)
Canu	1.7	Koren et al. (2017)
L_RNA_SCAFFOLDER	1.0	Xue et al. (2013)
SSPACE	1.0	Boetzer et al. (2011)
PBJelly	1.0	English et al. (2012)
Pilon	1.22	Walker et al. (2014)
gatk	4.1.3.047	McKenna et al. (2010)
PLINK	1.9	Chang et al. (2015)
LoRDEC	0.9	Salmela and Rivals (2014)

All crayfish animals from the wild populations were collected in accordance to local fishery regulations.

Table 3: ***Procambarus virginalis* animals used for genetic authentication and sequencing.** The names of wild populations, origination and method of authentication are indicated. Samples were collected by previous and current lab members or collaborators. Sample from Czech Republic was provided from laboratory stock established from the likely source of populations in the Czech Republic.

Country	Population name	Source
Germany	Reilinger See	Gutekunst et al. (2018)
Germany	Singliser See	S. Tönges, (2017)
Germany	Moosweiher	Vogt et al. (2015)
Germany	Baggersee Epple	S. Tönges, (2017)
Germany	Krumme Lanke	S. Tönges, (2017)
Austria	Karlsbader Weiher	D. Latzer, (2019)
Czech Republic	Vodňany	Patoka et al. (2016)
Slovakia	Bojnice	A. Kouba, (2018)
Slovakia	Leopoldov	A. Kouba, (2018)
Slovakia	Bratislava	A. Kouba, (2018)
Hungary	Danube	A. Weiperth, (2018)
Romania	Băile Felix	Pârvulescu et al. (2017)
Ukraine	Dnipro	R. Novitsky, (2018)
Estonia	Narva	Ercoli et al. (2019)
Malta	Ghajn il-Papri	Deidun et al. (2018)
Madagascar	Ihosy	R. Andriantsoa, (2019)

2.2.2 Sample preparations, DNA extraction, library preparation

- ***De novo P. virginalis* genome assembly.** The genomic DNA extraction was based on SDS method with isopropanol precipitation and performed by Katharina Hanna. SMRT large-insert library preparation was performed by collaborators from DKFZ Genomics & Proteomics Core Facility (Heidelberg, Germany) following the recommended protocols by Pacific Biosciences. Libraries were generated from 1 to 5 µg sheared and concentrated DNA with an insert size target of approx. 10 kb.

- **Whole-Genome Sequencing.** The extraction and preparation of genomic DNA of other *Procambarus* specimens were performed by Katharina Hanna and Sina Tönges. The DNA was isolated and purified from abdominal muscular tissue using a Tissue Ruptor (Qiagen), followed by proteinase K digestion and isopropanol precipitation. The quality of isolated genomic DNA was assessed via agarose gel elec-

Table 4: ***Procambarus fallax* animals used for genetic authentication and sequencing.** The names of populations, locality and origination are indicated. Speciments were provided by collaboration partners Christopher E. Skelton (C.E.S.) and Nathan J. Dorn (N.J.D.). The second collection of specimens from Loxahatchee Impoundments was performed in 2019.

#	Population name	Country/State	Source	Method
1	Grand Bay Creek	US/Georgia	C.E.S., June 2018	PCR, WGS
2	Harris Creek	US/Georgia	C.E.S., April 2018	PCR, WGS
3	Gainesville	US/Florida	C.E.S., June 2018	PCR, WGS
4	Welaka	US/Florida	C.E.S., June 2018	PCR, WGS
5	Silver River	US/Florida	(Manteuffel-Ross et al., 2018)	PCR, WGS
6	Lake Woodward	US/Florida	N.J.D., March 2019	PCR, WGS
7	Gum Slough	US/Florida	C.E.S., June 2018	PCR, WGS
8	Orlando	US/Florida	C.E.S., March 2019	PCR, WGS
9	Snell Creek	US/Florida	C.E.S., March 2019	PCR, WGS
10	Cypress Creek	US/Florida	C.E.S., March 2019	PCR, WGS
11	Three Lakes	US/Florida	N.J.D., March 2019	PCR, WGS
12	Hurrah Creek	US/Florida	C.E.S., March 2019	PCR, WGS
13	Blue Cypress Lake	US/Florida	(Levy et al., 2017)	PCR, WGS
14	Kissimmee River	US/Florida	N.J.D., January 2018	PCR, WGS
15	Bud Slough	US/Florida	C.E.S., March 2019	PCR, WGS
16	Arcadia	US/Florida	C.E.S., March 2019	PCR, WGS
17	Loxahatchee Slough	US/Florida	N.J.D., April 2018	PCR, WGS
18	Lake Okeechobee	US/Florida	N.J.D., February 2018	PCR, WGS
19	Loxahatchee Impoundments	US/Florida	N.J.D., April 2018	PCR, WGS
20	Corkscrew Swamp	US/Florida	N.J.D., April 2018	PCR, WGS
21	Everglades	US/Florida	N.J.D., October 2017	PCR, WGS
22	Everglades	US/Florida	N.J.D., November 2017	PCR, WGS
23	Everglades	US/Florida	N.J.D., November 2017	PCR, WGS

trophoresis and/or via 2200 TapeStation (Agilent). The Illumina libraries for WGS were prepared by colleagues from DKFZ Genomics and Proteomics Core Facility according to standard protocol.

2.2.3 Genotyping

Experimental part of genotyping held by Katharina Hanna and Sina Tönges. PCR genotyping of *P. virginalis* specimens from various European and Malagasy populations was performed using the mitochondrial and nuclear markers. The sequence with 277 nt of Cytb marker gene was amplified by using the following primer pairs: (FWD CAGGACGTGCTCCGATTCATG and REV GACCCAGATAACTT CATCCCAG). The sequence of 334 nt of Dnmt1 gene was amplified by FWD GCTTTCTGGTCTCGTATGGTG and REV CTGCACACAGCCTAAGATGC primers.

For *P. fallax* and *P. alleni* specimens apart from the previous markers, the additional primer pair for 553 nt sequence of Cox1 gene (FWD CTGCTATTGCTCATGCAGGT and REV TGCCCGAGTATCTACATCCA) was used. Selection of mitochondrial Cox1 marker was based on the initial analysis of marker gene sequence, which is identical for all *P. virginalis* and shows the stable single polymorphisms in *P. fallax* Cox1 gene sequence.

Three pair of primers were designed by Dr. Vitor Coutinho Carneiro.

Amplicons were verified by agarose gel electrophoresis and cloned using the TOPO TA Cloning Kit (Invitrogen) according to the manufacturer's instructions.

2.2.4 High-throughput sequencing

- For genotyping of *Procambarus* specimens from various populations (Table 3, 4) the purified plasmids were sequenced by Eurofins Genomics.
- SMRT sequencing for genome assembly was performed on a PacBio SEQUEL I platform according to the manufacturer's recommendations. Movie time ranged from 240 to 900, and average 600 for the most of SMRT cells. Sequencing a total of 37 SMRT cells resulted in 69,074,290 reads comprising 242,146,920,794 bases and an average read length of 6kbp.
- Whole genome sequencing was performed after genotyping for 15 *Procambarus virginalis* and 28 *Procambarus fallax* samples on the Illumina HiSeq platform by

paired end 150 bp protocol at DKFZ Genomics and Proteomics Core Facility. The high number of read pairs was obtained which allows to reach substantial coverage. The results are reported in the sequencing overview tables (Results Section, 7, 4). The raw data for 15 specimens of *Procambarus virginalis* has been deposited at the NCBI Sequence Read Archive under NCBI BioProject PRJNA599283.

2.3 Bioinformatic analyses

2.3.1 Analysis of genotyping

Amplicon sequences were aligned to *P. virginalis* reference gene sequences and compared using SnapGene software. The identified SNPs within each marker sequence were calculated and visualised by heatmap plot.

2.3.2 Genome assembly strategy

The *de novo* genome assembly based on application of SMRT long read sequencing dataset. In the first step of genome assembly workflow the Circular Consensus Sequences (CCS) were generated with the parameters: 0 minimum full passes and 70% minimum predicted accuracy (Figure 7).

By using publicly available highly accurate Illumina reads from Gutekunst et al. (2018), the long PacBio CCS reads were subjected to first error correction step. Extracted CCS reads were error corrected using LoRDEC software, which builds de Bruijn graph using only short reads. With generated path it searches for corresponding sequence within long read and in corresponding step corrects it. For this step the lordec-correct command was used with the following parameters for the sensitivity of correction:

- -k 19 for k-mer size =19, which is log4(the genome length).
- -s 3 for abundance threshold = 3 for the number of occurrences of a k-mer in short reads.

The second error correction step was based on long-reads self-correction within first module of Canu assembler. The command -correct was executed with the default parameters and specification on -pacbio-corrected input data.

After the correction 35,314,439 of long reads were generated and used for the main and the most time-consuming iterative assembly process. Canu software with the default parameters was used in the main genome assembly step.

After the iterative assembly the workflow includes the sequential application of scaffolding and polishing steps (Figure 7). The first scaffolding was performed by available long jumping distance sequencing dataset from Gutekunst et al. (2018) and aided by RNA-sequencing (RNA-seq) dataset using SSPACE-standard and L_RNA_SCAFFOLDER tools with default parameters (Table 2, Figure 7). SSPACE-standard scaffolder performs extension pf pre-assembled contigs via extraction of single reads and mapping onto contigs repeatedly. As a result contigs become jointed into longer scaffold sequences by introducing sequences with unknown bases (Ns). L_RNA_SCAFFOLDER scaffolder apply search of transcripts within genomic fragments and generates the path from optimal connections. This allows to extend scaffold length in the genome assembly.

The raw PacBio reads from the initial step were utilized for automated polishing by Pilon and gap filling by using PBJelly softwares. Pilon algorithm detects misassemblies and might initiate the new gap opening within the scaffold. The other gaps, single base errors and indels substantially eliminated from the assembly by Pilon. Additionally, computationally high-intensive PBJelly software performs the final gap closing.

On the final stage of assembly the second scaffolding was performed by proximity ligation based on Chicago and Hi-C methods provided by Dovetail Genomics ™ (Chicago, USA).

2.3.3 Genomic analysis

Trimming and pre-processing

The whole genome sequencing data was first quality assured using FastQC (Andrews, 2010) report. The raw reads were trimmed for adapters and low quality bases using Trimmomatic (Bolger et al., 2014) with the following parameters:

- ILLUMINACLIP: Remove adapters provided in a separate file. Trimmomatic looks for seed matches (16 bases) and allows a maximum of 2 mismatches. Seeds will be extended and clipped if a score of 30 is reached (paired-end reads), or of 10 (single-end reads).

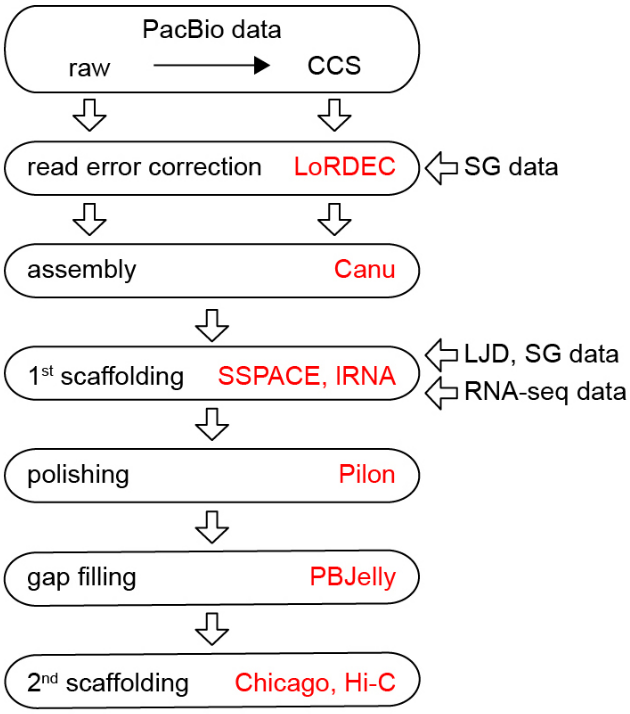

Figure 7: **Workflow of the marbled crayfish *de novo* genome assembly.** Seven major steps include: (1) the acquisition of raw long PacBio reads and generation of Circular Consensus Sequences (CCS); (2) the first long read error correction is performed with application of highly accurate Illumina read from Shotgun (SG) libraries; (3) the third step includes the second error correction step and iterative assembly process; (4) the forth step includes the first scaffolding step by utilizing Long Jumping Distance (LJD) and RNA-sequencing (RNA-seq) datasets; the final steps include (5) the assembly polishing, (6) gap filling and (7) second scaffolding by using Pilon, PBJelly and Chicago and Hi-C scaffolding provided by Dovetail™.

- LEADING, TRAILING: remove leading and trailing low quality or bases (below quality 3).
- SLIDINGWINDOW: for sliding 4-base windows, cut the read if an average quality of a window drops lower than 20.
- MINLEN: keep the reads with a minimum length of 36.

Mapping and subsequent read processing

Trimmed and pre-processed by removing duplicates reads were mapped onto:

- the *P. virginalis* mitochondrial genome sequence (Genbank KT074364.1),
- the V.04 draft genome,
- newly generated and reported as refined version of marbled crayfish genome assembly, V1.0 using bowtie2 with default parameters.

Resources of 10 cores with 20 threads and 100 MB RAM were allocated on the HPC cluster for mapping. Mapped reads were additionally committed to duplicate removal for reads pairs having identical external coordinates to retain the pair with highest mapping quality by using samtools rmdup function.

The alignments against the short contigs (shorter than 10 kb) of nuclear genome assembly were discarded to avoid loss of time in processing fragmented sequences using SAMtools. Alignments on mitochondrial genome were directly proceed for a pairwise alignment comparison.

Single Nucleotide Variant calling

The Bayesian genetic variant detector Freebayes (Table 2) was used for calling the variants against *P. virginalis* reference genome with the following parameters:

- report-all-haplotype-alleles, the command requesting the information about all alleles at sites where several haplotype alleles are detected.
- pvar = 0.7, to report the sites with the probability if there is a polymorphism is greater than 0.7
- ploidity = 3, adjusted for triploid genotype of *P. virginalis* and ploidity = 2 for diploid *P. fallax*.
- min-mapping-quality = 30, for excluding the alignments with the mapping quality less than 30.
- min-base-quality = 20, for excluding the alleles with the base quality lower than 20.
- min-coverage 6, to filter the sites having at least coverage of 6.

- report-genotype-likelihood-max, to provide the genotypes information using the maximum-likelihood estimate.

Variant file filtering

Variant filtering was performed on the set of multi-sample variant files and included filtering for read depth of 20, a maximum read depth of 200, and a minimum Phred-scaled quality score of 30 for each sample. For this the software gatk-4.1.3.047 was used (Table 2). Filtering for indels was performed by custom filtering pipelines.

Polymorphic site calculation in *P. virginalis*

The polymorphic sites of *P. virginalis* individuals were extracted from WGS dataset. The WGS dataset from the oldest known aquarium lineage sample (Heidelberg 1 (HD1), Gutekunst et al. (2018)) was used in comparative analysis as a reference. Only the sites covered in the reference genome were included in the comparative analysis. The natural heterozygous sites which are present in the reference genome were eliminated to narrow down the calculation of appearing polymorphic sites in non-heterozygous loci. Polymorphic sites for analyzed samples were identified by the presence of 1 nucleotide substitution to the reference sequence. All the other multiple substitutions were eliminated from the analysis.

After the multi-step filtering, the calculation of polymorphic sites was based on additive estimation of Single Nuclear Variations (SNVs) per sample in the common sites of variants in the multi-sample variant file.

Population structure analysis of *P. fallax*

Population structure was estimated by application of Principal Component Analysis (PCA) of identified variant positions. To avoid spurious correlations among the measured variables and keep the data independent for PCA, the genomic data was preprocessed. To avoid the impact of linkage disequilibrium on variant dataset, the set of identified SNVs was subjected to linkage pruning using PLINK (Chang et al., 2015) with the following parameter settings:

- allow-extra-chr, to get access to the additional chromosomes beyond the default human chromosome set.

- set-missing-var-ids @:#, for keeping the variants which have not been assigned by standard IDs.
- indep-pairwise 50 10 0.1, performs the linkage pruning with a set a window of 50 Kb, the window step size of 10 and r2 threshold of 0.1.

The resulting set of variants was used for a PCA using the default parameter settings in PLINK. The PCA results were plotted using ggplot2 package in R (version 3.2.1).

Population structure analysis of *P. virginalis*

Population structure analysis of *P. virginalis* individuals was performed after the initial preprocessing of dataset. First, the genotypes in variant file were re-coded into matrix with integers from 0 (0/0/0) to 3 (1/1/1) according to the alternative allele dosage. Finally, the PCA for each set of re-coded genotypes was run using the PCAMETHODS (Stacklies et al. (2007)) R package. Visualization of clusters was performed using Pheatmap package in R with default parameters.

Phylogenetic analysis

Phylogenetic analysis was performed based on nuclear whole-genome sequences of *P. virginalis* from diverse populations in Europe and on complete mitochondrial genome sequence for *P. fallax* individuals from populations in Florida and Georgia.

The generated matrix for detected nuclear variants was used to calculate the distances between pairs of *P. virginalis* individuals. The distance matrix was used for neighbour-joining tree estimation using the Ape package in R with the default parameters. The unrooted phylogenetic tree was generated using the Phytools package (Revell, 2012) with the default parameters.

The phylogenetic analysis based on *P. fallax* mitochondrial genome was performed in the context of haplotype analysis. Alignments after the mapping on *P. virginalis* complete mitochondrial genome (under the accession number KT074365.1) were used for SNP and consensus calling using SAMtools and BCFtools packages. Obtained consensus sequences of *P. fallax* mitochondrial genomes were subjected to the pairwise comparisons with the reference genome. The calculation of maximum distance matrix and phylogenetic tree generation were performed according to the pipeline described above in this section of methods.

Analysis of loss of heterozygosity

To analyse the heterozygosity the initial dataset with variants was used to select all heterozygous positions identified in reference genome sequence. The filtering criteria included selection of sites with Phred-scaled quality score >30 and number of reference observations >5 in the multi-sample variant dataset. Filtering resulted in a dataset with 250,453 heterozygous positions which were considered for Variant Allele Frequency (VAF) calculation. VAF per site was calculated as the ratio of the alternative allele count to the total number of reads covering the site of interest. The average VAF for each 10,000 positions was calculated for visualization of genome genome-wide variant allele frequency destitution. The same calculations were performed for ten genomes of *P. virginalis* representing independent populations.

2.3.4 Analysis of *P. virginalis* population in lake Reilingen

For the comprehensive analysis of parthenogenic marbled crayfish population the estimation of population size and population growth kinetic were encompassed. The estimation of the population size and population genetic variability calculation allowed to build the model of population dynamic of marbled crayfish population in Reilingen.

For the population size estimation, the mark-recapture method was applied. The sampling event at Lake Reilingen (49.296893N, 8.544591E) was performed in July 2018 by volunteer initiatives of the former and current Epigenetics group members under the guidance of Sina Tönges and Ranja Andriantsoa. Each sampling event included the capture, examination for previous marks, the next marking and release the animal back to the catch site. In total, 18 traps were allocated at different water locations and depths, with an average distance of 105 m between traps. Within 10 consecutive days the mark-recapture was performed in order to avoid any deviations causing by animal reproduction, death, migration and moulting.

The estimation of population density was based on Schnabel's method as an extension of the weighted average of Petersen estimates (Krebs, 2016) and performed by Ranja

Andriantsoa and Dr. Carine Legrand.

The simulation of population growth was based on three types of population growth model: exponential, logistic and growth based on Allee effect (Johnson et al., 2019). To simulate marbled crayfish population growth kinetics, the observed population estimates (population size) as well as growth parameters such as a growth rate of 200 (±100) offspring per animal and per year were used.

- The classical exponential growth model describes a population size at any time point t by (1) and the population size (2)

  $$N\left(t\right)=N\left(t_o\right)\exp^{g(t-t_0)} \tag{1}$$

  $$whN\left(i+1\right)=N\left(i\right)\exp^{g(t_{(i+1)}-t_i)} \tag{2}$$

  by initializing N(t_0)=1 for a clonal population with a growth rate of (exp(g)=200±100) per year.

- The logistic population growth model is described by (3)

  $$N\left(i+1\right)=N_i+gN_i\left(1-\frac{N_i}{N_{max}}\right) \tag{3}$$

  and the population size at time point t_{i+1}- t_i, where the maximum number of individuals was set to N_{max}: the current population size.

- The growth with strong Allee effect described by (4), where parameter A was chosen according to an established Allee threshold of -3.1 (Johnson et al., 2019).

  $$\Delta N=N_{i+1}-N_i=N_ig(1-\frac{A}{N_i}) \tag{4}$$

For calculating the probabilistic genetic variability of the Lake Reilingen population with the consideration of three different growth models, several parameters were used: a mutation rate of 3.6×10^{-9}(Liu et al., 2017) which is considered as a common for mutation rate in Arthropods, the genome size of 3,511,656,756 bp for marbled crayfish genome assembly (Gutekunst et al., 2018).

The observed genetic variability was calculated based on multiplication of number of identified SNVs from WGS dataset and the population size estimates. Observed and probabilistic genetic variabilities were compared for finding the intersection in the calculated values.

3 Results

In this chapter, the progress and results of each stage of marbled crayfish genome assembly are described in details. Refined genome promoted the detailed genomic analysis for search of marbled crayfish origination and genome variation analysis. The genetic variability of monoclonal marbled crayfish allowed the understanding of genome evolution and estimating of population growth model of clonal species.

3.1 Genome Assembly

The currently available draft of marbled crayfish genome was build by highly accurate short Illumina reads and thus resulted in generation of genome assembly consisting of short and fragmented scaffolds. This draft assembly covered approx. 50 % of the genome sequence and due to the presence of short scaffolds and contigs it represents unphased fragmented mosaic genome sequence. With the application of single molecule real-time (SMRT) sequencing technology provided by PacBio, the genome sequence was substantially improved by lowering the percentage of gaps and number of sequences with unknown bases (Ns). As a result, the average length of scaffolds in the assembly was considerably bigger.

The integrated genome assembly workflow consisted of seven major steps (Figure 7, described in the Materials and Methods section). The sequencing on 37 SMRT cells resulted in production of 69 million reads comprising 242,146,920,794 bases, which corresponds to 70X genome coverage. The average SMRT read length was 6 kb. According to the the genome assembly workflow, the raw continuous long reads (CLR) characterised by high error rate were subjected to the adaptor removal for generation of highly accurate Circular Consensus Sequences (CCS). As a result, more than 6.7 million subreads were obtained, which correspond to 8X genome coverage (Supplementary Table 10).

After the generation of raw and CCS reads the errors in the long reads are not completely eliminated. The present indels in the long reads can extensively impede the downstream analysis (Koren et al., 2012). Thus, the additional step of error correction is always considered in the *de novo* genome assembly workflow. The initial error correction step was performed by mapping the available short Illumina reads used in the previous draft genome assembly (Gutekunst et al., 2018). The Illumina sequences were generated from Shotgun (SG) and Long Jumping Distance (LJD) libraries with the raw yield of 245,157 Mbp.

The subsequent error corrections were performed within the assembly step using the internal error correction algorithm from Canu (Koren et al., 2017). The paramenters such as rawError-Rate and CorrectedError-Rate were set at 0.300 and 0.052 respectively. These parameters are adjasted for PacBio reads and define the allowed difference in the overlap between two uncorrected and between two corrected reads respectively. The following step after the error corrections was the central iterative assembly process performed by Canu assembler, which resulted in the initial *de novo* assembly consisting of 126,904 contigs (Table 5). The contigs are the continuous DNA sequences which are missing unknown bases in its composition. Thus, the initial genome assembly consisting of contings comprises of 0 of total number of N's in the sequence (Table 5, PacBio CCS assembly).

After obtaining the contigs, the genome assembly workflow follows to scaffolding step. The scaffolds are the joint contigs with introduced unknown sequences (Ns), which are inserted for the approximation of the scaffold length. The first scaffolding step in the marbled crayfish genome assembly workflow was performed with utilization of various softwares, where the available Illumina longer range mate-pair (Long Jumping Distance (LJD) and RNA-sequencing (RNA-seq) datasets were considered. IRNA software was used for performing the first scaffolding of obtained contigs and SSPACE-standard software was applied for the extraction of inserts information from mate-pair libraries ranging from 3kb to 20kb. These approaches increased the main genome scaffold sequence length from 2,911.410Mb to 3,568.271 Mb (Table 5), PacBio IRNA and PacBio SSPACE-standard). From this intermediate step, the genome assembly was further subjected to polishing and in subsequent steps, to gap-filling with utilization of initial raw CLR PacBio reads. The final step of the assembly workflow was the extension and increase in the number of scaffolds. Scaffolding was performed according to Chicago and Hi-C protocol provided by Dovetail™.

Table 5: **Genome quality metrics comparison of stepwise assembly results.** The quality parameters of the marbled crayfish draft genome assembly are shown in the PV04 column (1) (Gutekunst et al., 2018). The rest of the columns are referring to each step of the assembly workflow (Figure 7): (2) *PacBio CCS* is the result of the initial CCS-read-based genome assembly. (3) *PacBio IRNA* is the assembly after the first scaffolding step with application of transcriptomic data. (4) *PacBio SSPACE-standard* is the result of first scaffolding after utilization of Ilumina LJD dataset. (5) *PacBio SSPACE-standard Pilon* represents the assembly after the first scaffolding and polishing by using Pilon software. (6) *PacBio SSPACE-standard Pilon Dovetail* is the final refined version of genome assembly after the second scaffolding by Hi-C and Chicago methods provided by Dovetail Genomics ™

	PV 04	**PacBio CCS**	**PacBio IRNA**
Main genome scaffold total	3,394,710	0	375,027
Main genome contig total	3,784,331	126,904	381,796
Main genome scaffold sequence total (Mb)	3,511.657	1,041.445	2,911.41
Main genome contig sequence total (Mb)	1,848.994	1,041.445	2,910.729
Gap (%)	47.347	0.000	0.023
Main genome scaffold N/L50 (kb)	2,8228/29.652	3,1239/10.33	8,7390/9.652
Main genome contig N/L50	406069/932	31239/10.33	92877/9.511
Max scaffold length (kb)	717.999	99.426	327.391
Max contig length (kb)	70.296	99.426	219.314
Number of scaffolds >50 kb	11,742	131	1,496
% main genome in scaffolds >50 kb	30.8	0.74	3.60
GC	43.31	44.32	44.32
Total number of N's	1,662,662,583	0	680,579
	PacBio SSPACE-standard	**PacBio SSPACE-standard Pilon**	**PacBio SSPACE-standard Pilon Dovetail**
Main genome scaffold total	239,845	236,690	169,518
Main genome contig total	332,661	329,154	337,953
Main genome scaffold sequence total (Mb)	3,568.271	3,513.780	3,700.958
Main genome contig sequence total (Mb)	2,910.719	2,857.089	3,037.518
Gap (%)	18.428	18.689	17.926
Main genome scaffold N/L50 (kb)	1,7887/51.076	1,7370/52.277	983/144.428
Main genome contig N/L50	63585/12.025	61236/12.327	66419/12.234
Max scaffold length	784.599 kb	785.249 kb	73.606 Mb
Max contig length (kb)	227.967	227.922	213.11
Number of scaffolds >50 kb	18,380	18,372	8,152
% main genome in scaffolds >50 kb	50.70	51.46	64.98
GC	44.32	44.32	44.34
Total number of N's	657,552,299	656,690,120	663,440,479

More than 65% of the genome was assembled in scaffolds with the length longer than 50 kb. The final assembly comprises 18 % of missing (Ns) nucleotides which is comparable with the quality of recently reported complete genome assemblies of other organisms.

The genome assembly statistics for each step of genome assembly are described in Table 5.

3.1.1 Genome quality

One of the major quantitative metrics of genome assembly quality is the weighted mean sequence length (N50), which was increased from 9.6 kb to 144.4 kb in the final stage of the assembly workflow. N50 is the minimum length of scaffold or contig which covers 50% of the genome. Thus, the marbled crayfish refined genome assembly the scaffolds which are longer than 144.4 Kb and cover the half of the genome sequence length. In comparison to the previous marbled crayfish genome assembly, N50 was increased in 5 times (Gutekunst et al., 2018). The refined genome assembly reduced the number of total entries from more than 3 million to 170 thousand, and the number of unknown bases from 1,663 million to 663 million (Table 5).

3.2 Genetic relationships between *P. virginalis* and *P. fallax*

3.2.1 Selection *P. fallax* specimens

The selection of specimens for comparative genomic analysis was performed after genotyping the candidate samples which were obtained from diverse sites of *P. fallax* natural habitat. Genotyping of 92 *P. fallax* samples resulted in the selection of 1-2 representative specimens per each location for the whole-genome sequencing (Figure 8). Each of the selected specimens has at least one polymorphism within the sequence of marker gene, which verifies the absence of *P. virginalis* individuals among the sampled specimens from natural habitat of *P. fallax*. Additionally, the specimens of genetically distant crayfish species *P. alleni* (N=4) were included in the amplicon analysis as a control group, due to distinctively higher number of polymorphism within the marker genes (Figure 8). Thus, the highest number of SNPs was in the all three marker gene sequences was for *P. alleni* specimens. The *P. fallax* specimens with the lowest number of polymorphisms were selected

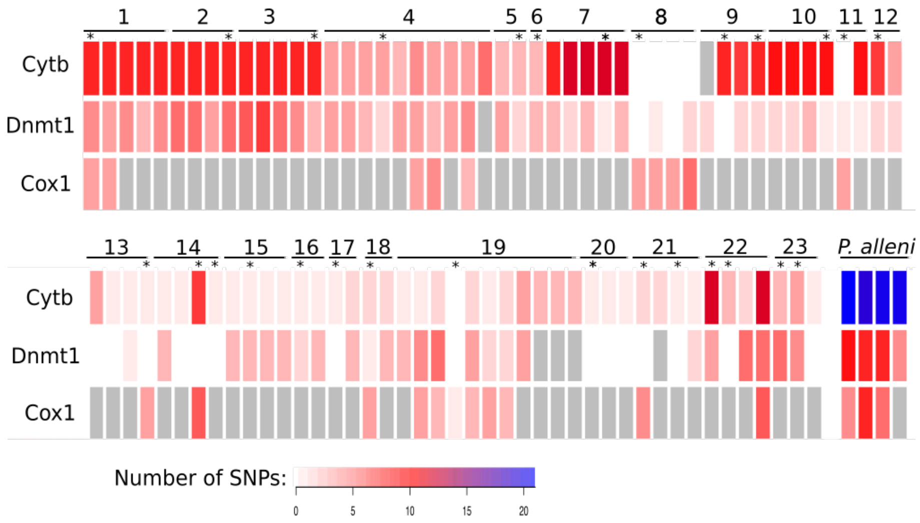

Figure 8: **PCR genotyping for three marker gene sequences.** The heatmap shows the number of genetic polymorphisms within each marker gene sequences in comparison with *P. virginalis* reference. The color scale range from 0 to 21. The distant species *Procambarus alleni* were used as control, which demonstrate the higher number of SNPs. The grey boxes indicate the missing information. The specimens pointed by asterisks were considered for whole-genome sequencing analysis.

for WGS (Figure 8).

3.2.2 Phylogenetic analysis

To perform phylogenetic analyses based on mitochondrial and nuclear genomes, the obtained reads were mapped on the the mitochondrial genome sequence reference (Genbank KT074364.1) and refined version of the marbled crayfish genome assembly (*P. virginalis*, V1.0) respectively. The mapping coverage ranged from 12.3 to 27.2. However, for the majority of samples the coverage was higher than 20, which was sufficient for in-depth genome-wide analyses (Table 6).

The phylogenetic comparison of mitochondrial genome sequences for representative individuals *P. fallax* with *P. virginalis* mitochondrial genome aimed to reveal the maternal haplotypes among *P. fallax* populations. The phylogenetic analysis resulted in observation of substantial genetic diversity of *P. fallax* mitochondrial genomes and identification of population closely related to *P. virginalis*. The diversity shown in the phylogenetic tree consists of well-defined branches, which suggest the existence of four major populations. These delineated populations were aligned with four major water catchment areas of Flor-

Table 6: ***Procambarus fallax* whole-genome sequencing overview.** 28 selected specimens for whole genome sequencing (WGS) resulted in sufficient genome coverage for further genomic analyses.

#	**name**	**sex**	**raw read pairs**	**processed read pairs**	**mapped (%)**	**covg. (X)**
1	Grand Bay Creek	f	423,783,854	337,745,407	82.9	27.2
2	Harris Creek	m	400,883,515	309,059,581	81.8	24.9
3	Gainesville	f	382,980,427	304,421,473	80.3	23.9
4	Welaka	m	419,925,777	338,428,570	82.3	26.7
5	Silver River	m	515,703,320	325,029,936	78.8	25.7
6	Lake Woodward	f	401,671,902	296,294,299	66.9	19.8
7	Gum Slough	m	429,448,611	339,413,271	80.4	26.7
8	Orlando	f	404,630,133	313,607,100	69.7	22.2
9.a	Snell Creek	m	405,889,320	182,146,721	64.2	12.3
9.b	Snell Creek	f	439,200,494	331,877,749	79.0	23.5
10	Cypress Creek	f	403,530,773	303,556,280	77.8	23.5
11	Three Lakes	f	401,945,296	310,359,901	81.0	24.4
12	Hurrah Creek	m	399,391,466	301,198,204	82.0	24.4
13	Blue Cypress Lake	m	428,390,609	332,127,247	82.5	25.4
14.a	Kissimmee River	f	408,765,829	324,407,550	83.4	25.9
14.b	Kissimmee River	f	438,654,151	356,033,853	78.2	27.0
15	Bud Slough	m	398,886,536	302,415,181	82.5	24.6
16	Arcadia	m	403,267,452	313,879,146	78.4	24.4
17	Loxahatchee Slough	f	419,442,289	208,892,712	85.2	18.1
18	Lake Okeechobee	m	388,988,492	299,392,496	81.4	23.4
19	Loxahatchee Imp.	f	390,255,144	308,958,034	86.5	25.8
20	Corkscrew Swamp	f	383,000,736	302,318,143	84.4	25.4
21.a	Everglades	f	427,850,688	336,421,204	82.8	26.8
21.b	Everglades	f	513,457,416	370,729,883	84.6	30.4
22.a	Everglades	n.d.	433,735,326	347,901,252	83.1	28.0
22.b	Everglades	n.d.	424,902,716	339,802,442	82.9	27.2
23.a	Everglades	n.d.	432,856,065	343,689,994	81.8	24.9
23.b	Everglades	n.d.	423,783,854	294,515,468	80.3	23.9

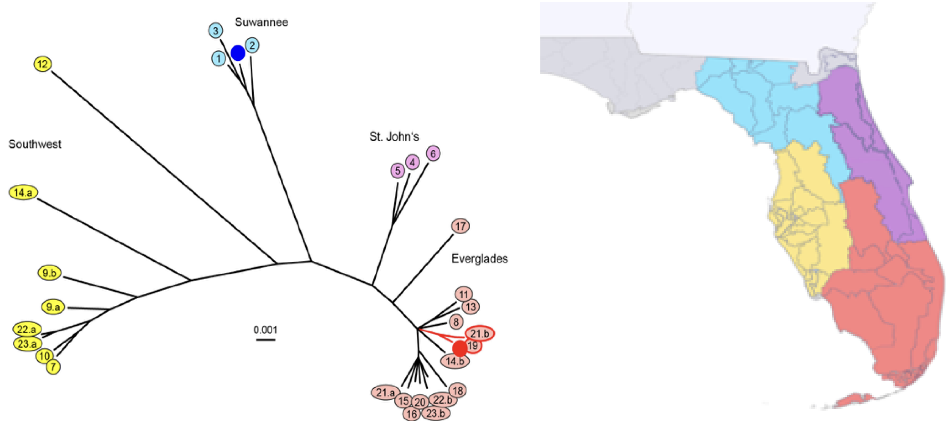

Figure 9: **Phylogenetic tree, based on complete mitochondrial genome sequences and four major water catchment areas of Florida.** LEFT: Phylogenetic tree. Colors in the phylogenetic tree indicate phylogeographically defined populations that were named after the corresponding water catchment areas of Florida. The darkblue dot indicates the *P. fallax* mitochondrial genome reference sequence, the red dot indicates the monoclonal *P. virginalis* mitochondrial genome reference sequence. *P. fallax* specimens which are particularly closely related to *P. virginalis* are highlighted by red lines. RIGHT: The four major water catchment areas of Florida: Suwannee River (blue), St. John's River (purple), Southwest (yellow) and Everglades (red).

ida (Figure 9). Interestingly, that among defined groups of *P. fallax* populations, some specimens sampled from a single site were the representatives of two different populations (Southwest and Everglades). This observation illustrates the presence of sympatry in the populations (site Everglades, 22-23, marked as *a* and *b*). The *P. virginalis* mitochondrial genome used as the reference for comparative analysis was placed in the Everglades population, in a close relation to individuals from Everglades population (19 and 21.b). The number of SNPs identified within the complete mitochondrial genome sequences of these samples was 12 and 8 respectively (Figure 9). Noteworthy, that found two closely related to marbled crayfish specimens, 19 and 21.b were collected from independent locations in the south-eastern Everglades with 70 km distance.

Following phylogenetic relationship analysis for nuclear genome sequence was performed by Dr. Julian Gutekunst in collaborative work. A phylogenetic comparison was performed separately for both A and B haplotypes based on the previously established AA‘B haplotype structure (Gutekunst et al., 2018). The results showed a population structure that was largely similar to the structure observed in comparison of mitochondrial genomes. Therefore, similarity in phylogenetic comparisons suggests that both *P. virginalis* haplotypes are related to the *P. fallax* from Everglades population. The nuclear sequence comparisons for each haplotype (A and B) again showed the particularly close relationship of the two specimens from Everglades population (19 and 21.b) as most closely related to *P. virginalis*. These findings suggest that the marbled crayfish is a direct descendant of *P. fallax*, with both parental haplotypes inherited from the Everglades population.

3.2.3 Population structure analysis

The refined genome assembly was used for estimation of *P. fallax* population structure analysis. The population structure analysis of *P. fallax* species was based on comparisons of identified variant patterns within each sample from 28 specimens set. The variant patterns were obtained from subsequent alignment and variant calling against the reference genome. The generated set of sample-specific variants was investigated by Principal Component Analysis (PCA) for establishment of population structure. The results showed overall similar population structure to the shown in the phylogenetic analyses. It also indicates the heterogeneous structure in the Everglades and Southwest populations of *P. fallax* species and indicates the presence of similar genetic features between specimens 19 and 21.b (Figure 10).

3.3 Genetic relationships between *P. virginalis* populations

The clonality of the marbled crayfish genome was observed and first described in the previous study (Gutekunst et al., 2018), where a few hundred of SNVs were identified. These findings suggested the presence of low genetic variability in the monoclonal genome. It was assumed, that the main source of the identified genetic variability was the natural mutation accumulation within the genome of lineages over time (Gutekunst et al., 2018). In this PhD thesis, more detailed genome-wide analysis of SNPs was performed and the population-specific patterns of polymorphisms in marbled crayfish were analysed. These analyses were aimed to uncover the genetic relationship between the worldwide geographically distant marbled crayfish populations, as well as genetic variability within the single population.

In this subsection the *P. virginalis* sample selection is described. First, the initial genetic authentication of specimens from diverse and independent populations was performed

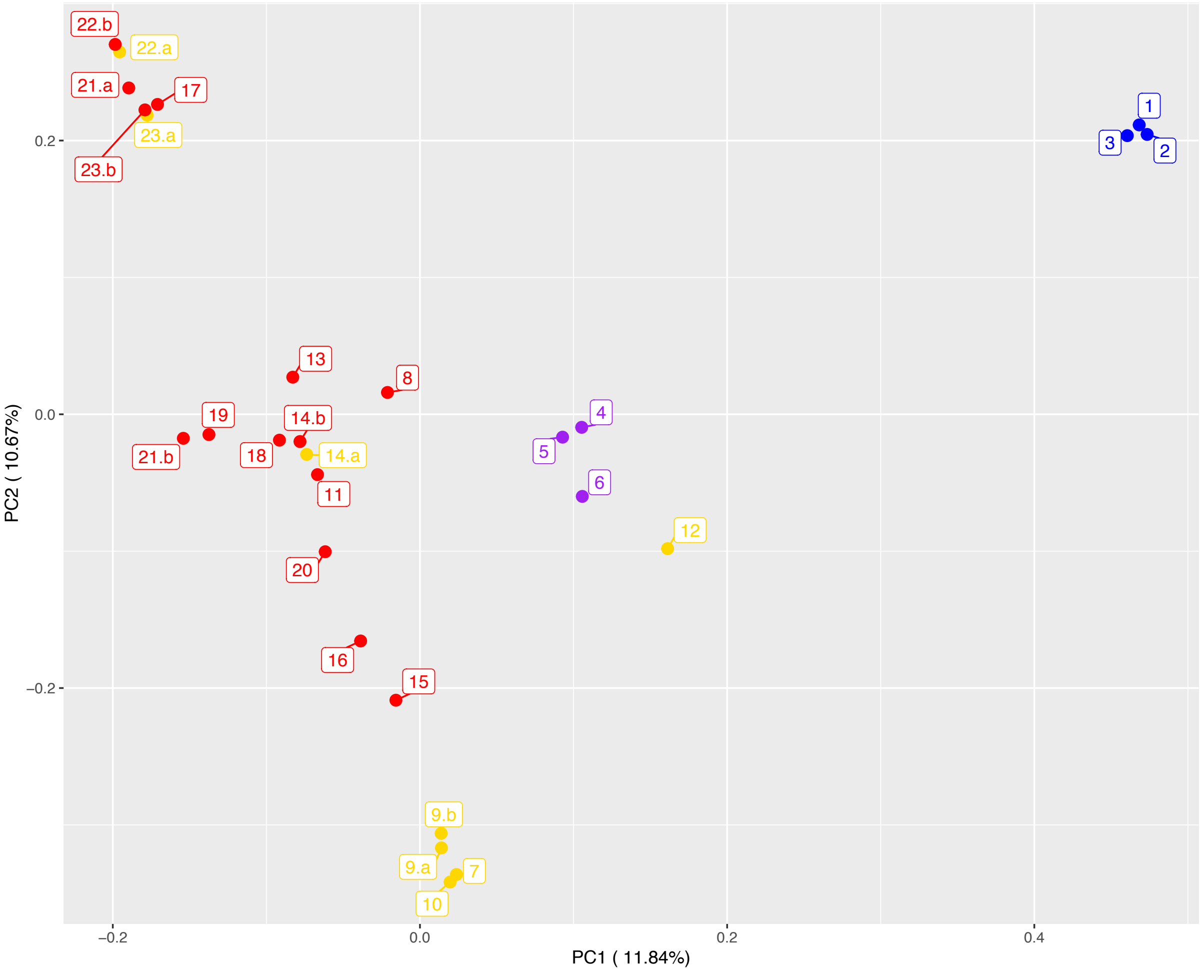

Figure 10: **Principal component analysis, based on the complete *P. fallax* WGS dataset.** Colors indicate phylogeographically defined populations: Suwannee River (blue), St. John's River (purple), Southwest (yellow) and Everglades (red). Three locations (14, 22 and 23) show the local sympatry of Southwest and Everglades populations

which followed by selection of each representative specimen from European and Malagasy populations for in-depth analysis of genomic variability by utilizing WGS.

3.3.1 Genetic authentication of *P. virginalis* populations

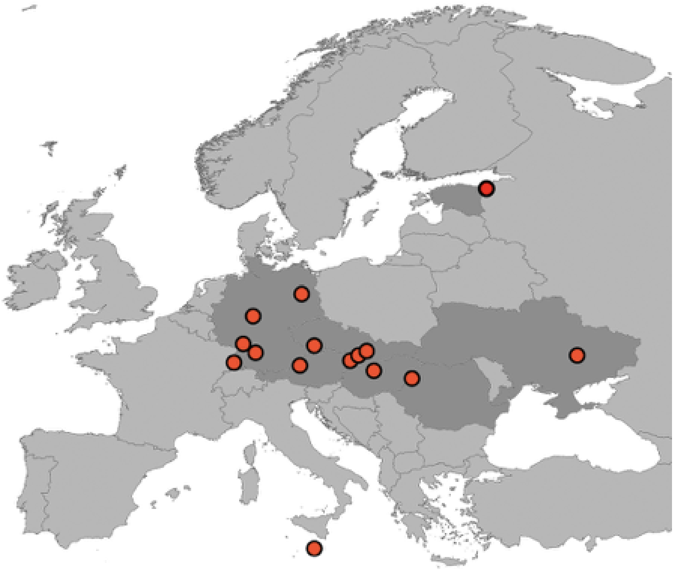

Figure 11: **Map of Europe and indicated locations of stable wild populations.** The countries with stable wild marbled crayfish populations are highlighted in dark grey. 15 population from nine European countries are indicated by red dots.

The various wild invasive populations of marbled crayfish were suggested according to the literature, where the recent records of marbled crayfish stable populations were reported (Jones et al., 2009; Chucholl, 2015; Lipták et al., 2016; Novitsky and Son, 2016; Patoka et al., 2016; Pârvulescu et al., 2017; Deidun et al., 2018; Gutekunst et al., 2018; Ercoli et al., 2019) (Figure 11). The other countries were also known for appearance of wild marbled crayfish populations, however these records were not confirmed in the recent years (Vojkovská et al., 2014; Bohman et al., 2013; Cvitanić et al., 2016). The possible origin of marbled crayfish populations can be the result of recent anthropogenic releases. The used specimens originate from the confirmed stable wild populations and were kindly provided by Ranja Andriantsoa, Sina Tönges, Lucian Parvulescu, Roman Novitsky, Alan Deidun, Fabio Ercoli and Antonin Kouba (Table 3). In total, the marbled crayfish specimens from 15 wild populations were considered for the genetic comparisons.

Table 7: ***Procambarus virginalis* whole-genome sequencing overview.** 15 samples selected for WGS include the specimens from 10 European populations and one from Madagascar

Population name	Country	Number of reads	Mapping ratio (%)	Coverage
Reilinger See	Germany	362,022,733	84.0	42.7X
Singliser See	Germany	431,347,724	82.0	39.6X
Moosweiher	Germany	420,501,584	82.9	40.2X
Bojnice	Slovakia	410,365,132	76.4	35.3X
Ghain il-Papri	Malta	412,900,336	82.7	38.6X
Vodňany	Czech Republic	425,171,972	81.6	39.0X
Danube	Hungary	408,137,880	77.2	36.4X
Băile Felix	Romania	446,242,415	81.7	41.5X
Dnipro	Ukraine	419,042,145	81.8	38.3X
Narva	Estonia	417,935,760	81.8	39.6X
Reilinger See (2)	Germany	452,534,084	81.5	40.6X
Reilinger See (3)	Germany	454,991,136	81.7	40.9X
Reilinger See (4)	Germany	451,122,510	81.3	40.7X
Reilinger See (5)	Germany	458,780,022	80.8	42.1X
Ihosy	Madagascar	452,157,698	80.3	36.8X

3.3.2 Identification of population-specific SNVs in *P. virginalis* genomes

For this purpose the WGS dataset was obtained for 10 animals, representing 10 independent populations from 8 European countries (Table 7). The additional dataset of four representative individuals from a single population inhabiting Reilingen Lake was obtained.

The average mapping coverage was higher than 30 for most of the samples which was sufficient for detailed genetic analysis (Table 7).

The comparative analysis of genetic variants included more than 300,000 genomic sites. After the filtering for quality and coverage the number of identified single nucleotide variants (SNVs) was considerably low and ranged from 5,722 to 7,537 among all the samples considered for population analysis. This is an order of magnitude below the number of SNVs observed in other clonal organisms, such as in daphniids (Muñoz et al., 2016). The low number of SNVs in marbled crayfish can be explained by young evolutionary age of species. Identified SNVs were used to determine the population-specific patterns. The genomes from ten European populations showed the homogeneous variability, reflected in low number of unique and high number of inter-sample shared variants (Figure 12). Thus, the number of identified SNVs that are shared by 6 to 9 animals were an order higher than the number of unique or partially shared by 2-5 samples. (Figure 12). The number of unique SNVs for each independent populations was noticeably low, suggesting the recent separation of lineages.

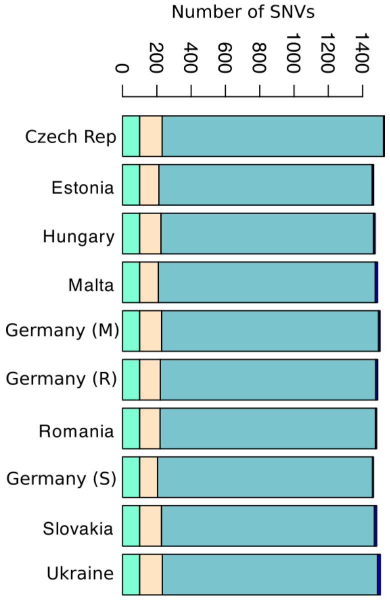

Figure 12: **Stacked bar plots, illustrating the number of unique and shared SNVs.** sd.2-5 indicates SNVs that are shared by 2 to 5 animals while sd.6-9 indicates SNVs that are shared by 6 to 9 animals.

To compare the genomic variations which occur within the single population, four additional animals representing a single population from Lake Reilingen in Germany were considered for the genomic variability analysis (Table 7). The WGS dataset was prepared and analyzed according to the established pipeline described in the Materials and Methods Section.

Among identified genome-wide single variations within samples (N=5) from Reilingen population, a set of 3,635 shared SNVs was determined. Among the identified Reilingen-specific SNVs 1076 SNVs were shared with SNVs identified in the diverse marbled crayfish European populations. The number of SNVs shared between 9 other European populations was 2,607 (Figure 13, A). The number of shared SNVs for European populations was lower due to the higher heterogeneity caused by higher number of samples considered for the comparative analysis. The detailed analysis of Reilingen-specific SNVs showed notably low sample-specific SNVs, demonstrating a negligible variability within population (Figure13, A).

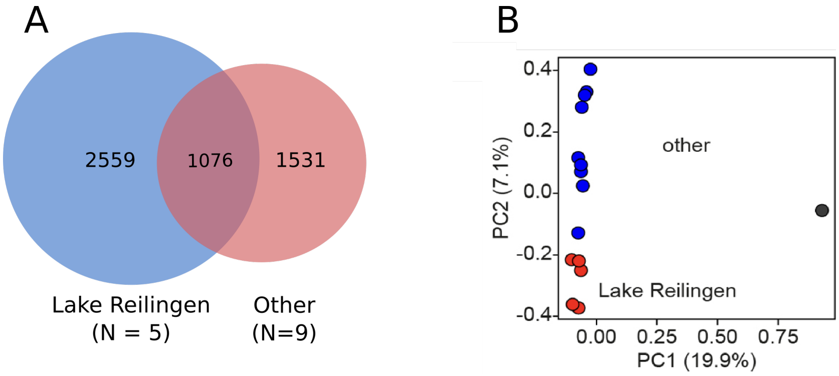

Figure 13: **Venn diagram and principal component analysis of shared SNVs between Reilingen population and 9 other European populations.** (A) Venn diagram indicating the level of overlap between SNVs identified in marbled crayfish genome from Reilingen population and other European populations. (B) PCA based on total 7,858 SNVs identified within the marbled crayfish genome from Reilingen and other European populations.

To identify the genetic population similarities and to perform the population structure analysis, the WGS dataset according to the alternative allele dosage was recoded into matrix (see Materials and Methods section) and used for principal component analysis. PCA was based on 7,858 total set of polymorphic sites, which include unique and shared SNVs between the populations. PCA result shows a clear segregation of European populations. Moreover, the similarity between SNV patterns within Reilingen population resulted in separation of Reilingen group with a close proximity to each other, showing particularly close genetic relationship within the population (Figure 13, B).

To investigate the genetic variations in geographically distant populations, the publicly available high-coverage WGS datasets of independent populations from Madagascar together with the dataset produced for the additional specimen of Ihosy population (Madagascar) were integrated for the analysis. After the data processing and variant calling for the common genomic sites, a set of 7,537 SNVs were identified for the European (N=10) and Malagasy (N=5) samples. A comparatively high number of shared SNVs for European and Malagasy populations likely reflects the distance of the wild populations from the ancestral aquarium lineage (Figure 14, A). PCA based common variant sites (7,537) from independent locations in Europe and Madagascar showed a segregation into two clusters according to the geographic region of origin. Despite the nigh number of shared genomic variants, the geographically distant populations can be separated into European and Malagasy groups (Figure 14, B).

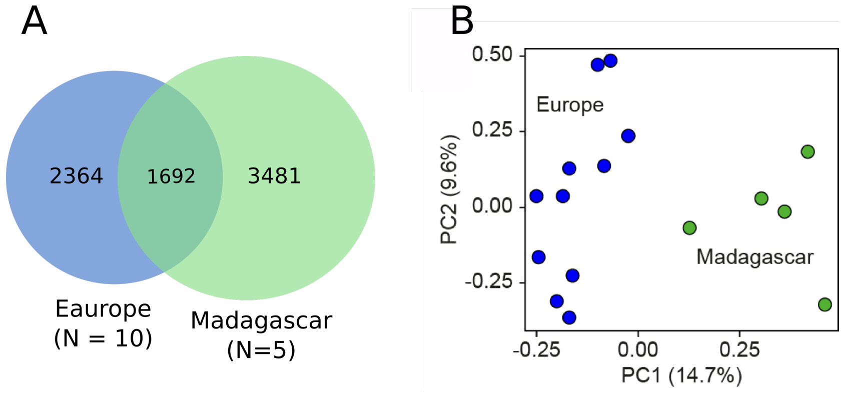

Figure 14: **Venn diagram and principal component analysis of shared SNVs between Malagasy and European populations.** (A) Venn diagram indicating the level of overlap between SNVs identified in marbled crayfish genome from 5 Malagasy and 10 European populations. (B) PCA based on total 7,537 SNVs identified within the marbled crayfish genome from Malagasy and European populations.

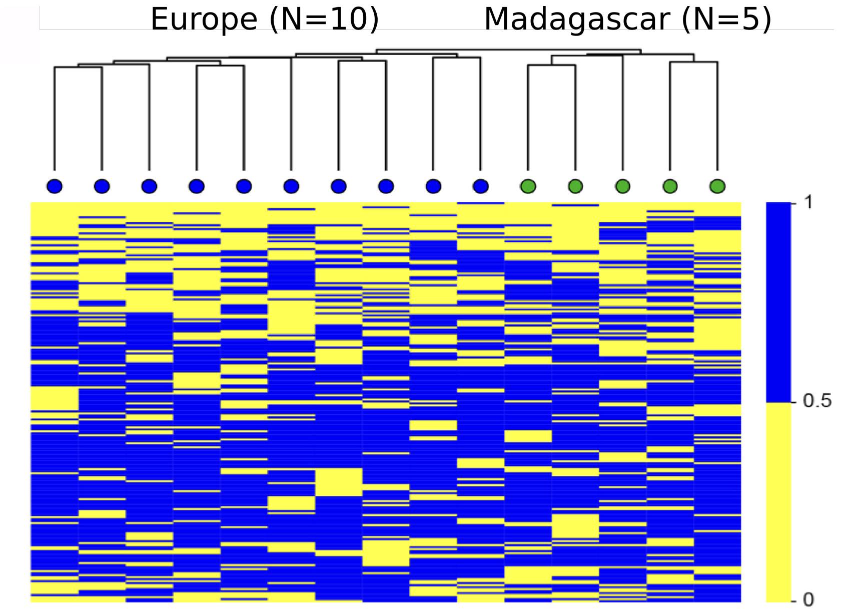

Figure 15: **Heatmap of the single variant patterns based on the 8,912 common polymorphic sites.** Clusterization was with default parameters and kmeans_k = 200 for aggregation the rows with variant sites. The average number of variants per 200 sites shown by the color scale. The clustering reflects the specific geographical SNV patterns: the European and Malagasy populations are indicated by blue and green dots correspondingly

Finally, to better visualize the SNV population-specific patterns, all analyzed samples were recoded into alternative allele dosage matrix for generating a heatmap with default clusteriation criteria (see Materials and Methods for details). The results showed the presence of segregation into several clusters, particularly aligned to the European and Malagasy populations (Figure 15), consistent with the previous observations.

The full set of identified shared and sample-unique SNVs from all 15 populations were annotated (N=16,564). The results show the majority (75%) of variants localized in the intergenic regions and only 4% localized in the coding regions (Table 8, Figure 16). Moreover, the analysis of base changes showed a high abundance of G to A and C to T base substitutions (Table 9) with transitions / transversions (Ts/Tv) ratio of 1.3. The variant rate within the effective genome length (172,036,703 bp) characterized by 1 variant per each sequence of 10,336 nt length. The similar patterns have also been observed in the other organisms among arthropods, including parthenogens like pea aphid and water flea (Fazalova and Nevado, 2020; Flynn et al., 2017).

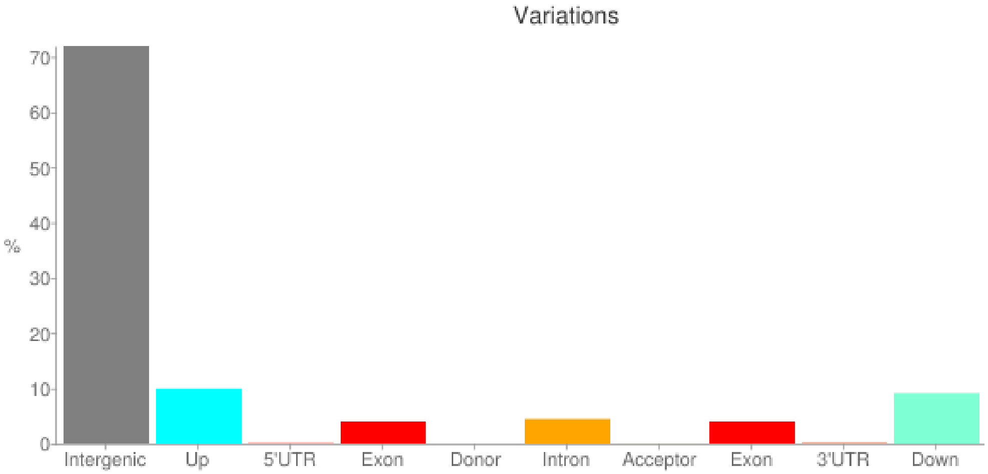

Figure 16: **Proportion of single nuclear variations in the genome region.** Percentage of the identified SNVs according to the genome region. The majority of variants are located in the intergenic region.

Identification and analysis of population-specific SNVs illustrate the emergence of genomic variability in the marbled crayfish genome from the geographically separated populations. The number of identified nuclear variations were remarkably low although sufficient to distinguish the lineage-specific mutational patterns.

Table 8: **Counts of SNVs classified according to variant effect types** The majority of variants were found in the intergenic region.

Type (alphabetical order)	Count	Fraction
3‘_UTR_variant	24	0.21%
5_prime_UTR_premature_start_codon_gain_variant	3	0.03%
5_prime_UTR_variant	13	0.13%
downstream_gene_variant	1,033	9.16%
intergenic_region	8,116	72.00%
intron_variant	501	4.44%
missense_variant	255	2.26%
missense_variant+splice_region_variant	4	0.04%
splice_acceptor_variant+intron_variant	1	0.01%
splice_region_variant	2	0.02%
splice_region_variant+intron_variant	6	0.05%
stop_gained	11	0.10%
stop_lost+splice_region_variant	1	0.01%
synonymous_variant	182	1.61%
upstream_gene_variant	1,121	9.94%

Table 9: **Nucleotide base changes caused by SNVs.** The high abundance of G to A and C to T base substitutions result in transitions/transversions ratio of 1.3.

	A	C	G	T
A	-	348	1,011	640
C	862	-	247	1,584
G	1,618	284	-	827
T	637	1,044	380	-

3.3.3 Phylogenetic analysis of *P. virginalis* populations

The observed genetic similarities between European marbled crayfish populations and their genomic homogeneity were demonstrated by phylogenetic distance tree. The phylogenetic analysis was based on generation of distance matrix of variations following by application of Neighbor-Joining method. Thus, the resulting tree shows the individual branches for marbled crayfish genomes from diverse European populations and longer distance towards the the ancestral aquarium lineage, indicating the the equivalent mutational activity in the genomes of current European populations. (Figure 17)

To clarify monoclonality of marbled crayfish genome, the phylogenetic analysis based on the complete mitochondrial genome sequence was performed. In this analysis the mitochondrial genome sequences of sexually reproducing parental species *P. fallax* from four independent populations were considered. The result showed the allocation of 11 *P. virginalis* specimens from the independent diverse European populations together with reference (HD1) on the one branch. The other *P. fallax* mitochondrial genomes showed substantial heterogeneity dividing into four branches which correspond to four major populations in Florida. This phylogenetic analysis confirms the monoclonality of marbled crayfish and demonstrates the high genetic identity of maternal genotype within parthenogenic *P. virginalis* species (Figure 18).

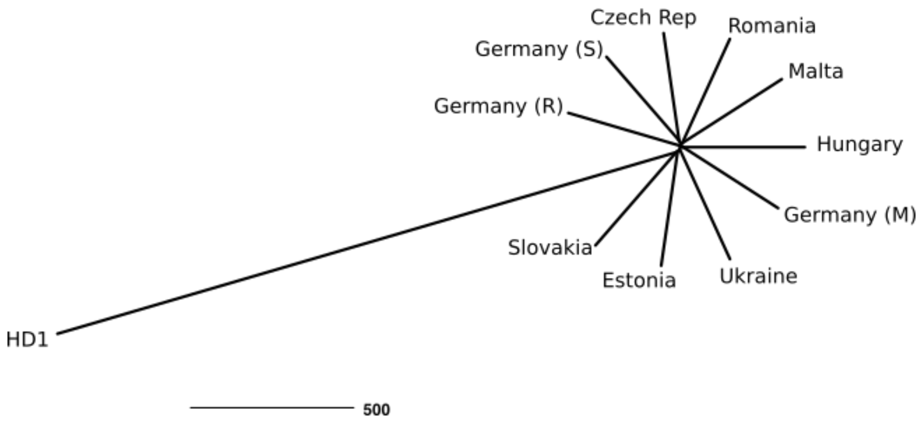

Figure 17: **Distance tree based on generated variation matrix** The distance counts were calculated from the genome variations identified within the genomes of the marbled crayfish from ten independent European populations. The scale bar: 500 counts. HD1 indicates the specimen of possible foundational aquarium lineage of marbled crayfish. Referred "Germany (R/S/M)" indicates German populations from Reilingen, Singliger See and Moosweier.

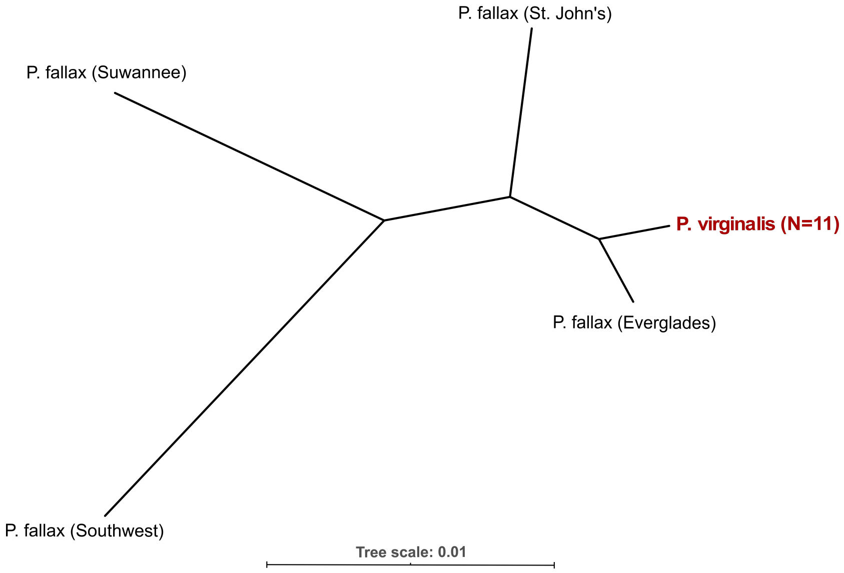

Figure 18: **Phylogenetic tree based on complete mitochondrial genome sequence.** 15 mitochondrial genomes of *Procambarus* animals were considered for the analysis. 11 *P. virginalis* specimens from independent European populations together with HD1 reference demonstrate monoclonality of mitochondrial genomes and 4 *P. fallax* from independent populations distributed in four branches. The scale bar: 0.01.

3.3.4 Test for loss of heterozygosity in the clonal marbled crayfish genomes

Long or short tracts of loss of heterozygosity (LOH) are the result of recombination in apomixis and prominent for asexually reproducing species, including well-studied *D. pulex* (Goudie et al., 2012; Flynn et al., 2017). LOH is one the the way to increase genetic variability at the absence of sexual reproduction. To investigate this genetic consequence of asexuality and the presence of ameiotic recombination in the marbled crayfish genome, the test for loss of heterozygosity was performed.

The genome-wide analysis of variant allele frequency showed the identical distribution of alternative allele frequencies in ten marbled crayfish genomes from independent populations (Figure 19). The evidence for loss of heterozygosity was not revealed, which is consistent with the suggested apomictic parthenogenesis in marbled crayfish. Thus, the detected genetic variation in marbled crayfish is assumed to be driven by the natural mutational process in the genome.

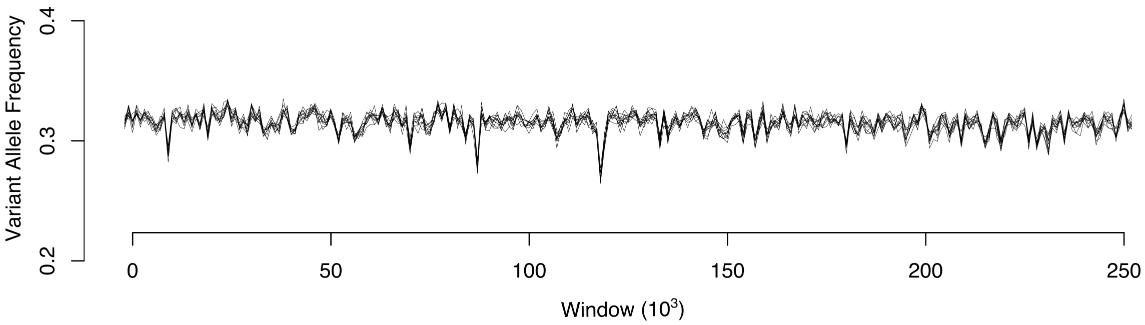

Figure 19: **Genome-wide analysis of variant allele frequency.** Lines represent the average variant allele frequencies per window of 10,000 heterozygous positions for each of the ten reference specimens. Variant allele frequencies are highly similar for all specimens and do not indicate a loss of heterozygosity.

3.4 Population analysis

The observed genomic variations in the clonal populations is an important foundation for analysis of variability in a growing population and for understanding the kinetics of population growth. One of the established marbled crayfish population in Reilingen Lake was subjected to the population analysis using the classical population estimation methods.

First, the current population size at Lake Reilingen was assessed by using the mark-recapture method and trapping. The calculation resulted in a population size estimate of 22,962 (SE=3,613) sexually mature animals >6 cm. With consideration of animals with smaller body size the total population size estimate resulted in 192,000 (SE=67,000) animals. The high deviation of the total population size estimate is largely due to the difficulty to accurately determine the number of small animals in the lake.

Second, the estimated total population size was used for the estimation of probabilistic total genetic variability occurring within the population for the inferences about the growth model of the population. Probabilistic genetic variability established for current population in Reilingen Lake based on WGS data and population size estimation was compared to those that might occur in populations under different growth models. Thus, logistic growth, exponential growth and growth based on Allee effect were used for simulation of the marbled crayfish population growth and calculation of corresponding genetic variability that occur in each of them (see Materials and Methods for details).

Among three growth models, exponential growth showed more representational result demonstrating the simulated total genetic variability in marbled crayfish population intersecting with the observed genetic variability of the current Lake Reilingen population at the time point 3.7 (±0.3) years (Figure 20). These findings suggest that marbled crayfish populations can have a very rapid initial growth phase and increase the population size according to exponential model. However, due to habitat resource limitations these clonal populations are expected to reach saturation, which was not observed in Reilinegen lake during the observation time.

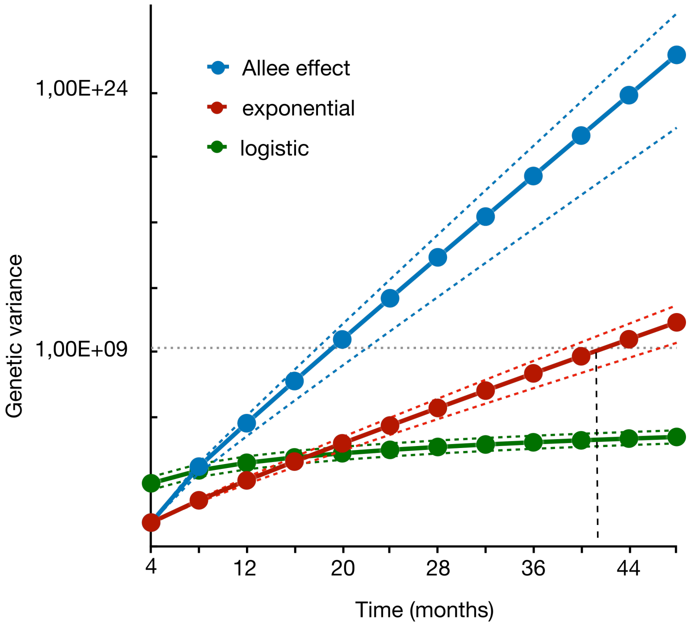

Figure 20: **Marbled crayfish population growth kinetics simulations.** The horizontal dashed line represents the genetic variability of 1e+09 in the current Reilingen population. Colored lines show the predicted genetic variability for exponential (red), logistic (green) and Allee (blue) growth models. Dashed colored lines indicate error margins.

4 Discussion

The cost cutting in sequencing and improved quality in assembly algorithms of short-read platforms resulted in substantial increase in the number of sequenced genomes (Schatz et al., 2010). However, most of the assemblies are highly fragmented and require an iterative refinement. Recently the first draft genome assembly of marbled crayfish was released, which was built using exclusively Illumina sequencing data (Gutekunst et al., 2018). As expected for short-read based assemblies provided by Illumina, the *de novo* draft assembly resulted in 50% of gap sequences, ambiguities and errors, which impeded more detailed downstream applications. Thus, the quality of the draft assembly disrupted the detection and annotation of genes, which apparently were underestimated in the initial automatic annotation (21,000 genes) (Gutekunst et al., 2018). Nevertheless, the draft assembly of marbled crayfish enabled the first in-depth phylogenetic comparisons with other *Procambarus* species and provided the first set of data pointing towards the origination of marbled crayfish from a single clone. The first detailed analysis of genome-wide heterozygous variant distribution showed unique biallelic and triploid (AA'B) genomic structure with high heterozygosity level. These findings also supported a first model for the origination of marbled crayfish by autopolyploidization, which was also suggested before (Vogt et al., 2015; Martin et al., 2016).

A unique genome structure, monoclonality and origination of marbled crayfish initiated the further studies, which required a refined version of genome assembly. Genome completeness is particularly important for genome-wide variant detection in comparative analysis of monoclonal species. This was possible only after introducing the third generation sequencing technologies and modernization of assembly algorithms.

This section covers the discussion of the results obtained for the refinement of marbled crayfish genome assembly and its important applications in the research of the genome origination, clonality and evolution.

4.1 An improved *Procambarus virginalis* genome assembly

For refinement of the marbled crayfish genome SMRT PacBio sequencing (Pacific Biosciences of California, Inc.) was chosen due to its numerous advantages which were already described in the Introduction section. However, Illumina sequencing data provided by original first draft assembly was used in order to correct highly erroneous long read. A multistage workflow was established, which included long read correction, assembly, scaffolding and further polishing steps. The final refined version of the marbled crayfish genome (version 1) reached 3.7 Gb size which substantially improved completeness. Only 18% of the genome consisted of unknown gap bases which are comparable with other well-curated reference genomes of important model organism, such as *Drosophila melanogaster* (Peona et al., 2018) or human (Chaisson et al., 2015a).

The newly build refined version of marbled crayfish genome allowed the detection of 133,654 gene models through automatic annotation. The increased number of gene models was identified due to the higher rate of genome completeness and application of more sensitive prediction parameters. Furthermore, the improved genome assembly also substantially expanded the coverage of repeat elements in the genome. Thus, it the annotation resulted in detection of 2,067,102 repeats, among which are SINE, LINE, Long terminal repeats (LTRs), simple repeats, low complexity regions and others. In total, repeat elements represent 14.8% of the annotated genome assembly. In the the genome of other arthropod species repeat elements range from 9% to 17% (Colbourne et al., 2011; Elsik et al., 2014; Sadd et al., 2015) and thus, suggest the comparable quality of the marbled crayfish reference genome.

The biggest challenge of assembling PacBio data was the absence of established pipelines for large-sized polyploid genome assembly. Due to novelty of long-read technology, the assembly iterative operation required the numerous trials to adjust the parameters and design the most advantageous workflow. Adjacent of parameter was essential for obtaining a higher quality assembly and for the efficient use of the computational resources. After several attempts in comprehensive error corrections, the final workflow took almost one year to complete the assembly and reach the final 3.7 Gb genome size. Although, the increase the size of computing resources would lead to reduction of assembly time.

4.2 How complete is the "complete" genome of marbled crayfish?

The highly curated genome assemblies usually named as a complete genome, including the assemblies of most popular model organisms. However, these genomes have the regions highly enriched in repetitive DNA, which are usually underrepresented in assemblies. With the availability of long-read sequencing technologies many genome assemblies were improved successfully characterizing many of these underrepresented repeat-rich regions. The refined version of the marbled crayfish genome (v.1) does not correspond to expected haploid chromosome number and thus, the genome is not complete. A truly "complete" assembly should contain the full set of chromosomes as a gap-free sequences, including the centromere and telomere regions. Nevertheless, the current marbled crayfish genome assembly is substantially improved in comparison with previous draft assembly and comparable with other genome assemblies of the model organisms. Still it subjected for curation, similar to one which is still happening for human genome assembly. In the latest version of human genome assembly no single chromosome has been finished end to end, and hundreds of unresolved gaps persist in it (Jain et al., 2018; Schneider et al., 2017). The main way of genome assembly curation is the long-range scaffolding, which includes linked-read cloud sequencing (Weisenfeld et al., 2017), nanochannel optical mapping (Seshadri et al., 2018) and chromosome conformation capture, Hi-C (Dudchenko et al., 2017). These scaffolding methods allow to achieve chromosome-sized scaffolds with minimized gap sequences. Chicago and Hi-C protocols provided by Dovetail Genomics ™ were applied for curation of marbled crayfish genome assembly, which substantially improved contiguity and accuracy of the assembly, but did not reach chromosome level. The biggest challenge in the genome construction is to assemble and anchor assembled sequences to chromosomes, particularly in construction of autopolyploid and highly heterozygous genomes.

4.3 Marbled crayfish genome origination

The newly generated marbled crayfish reference genome opens up on a lot of research possibilities. First of all, it allows a detailed analysis of genetic relationship between *P. virginalis* and its parental species *P. fallax*. Phylogenetic analysis based on mitochondrial and nuclear genomes sequences resulted in identification of both parental haplotypes in the native *P. fallax* population inhabiting in the Everglades region of tropical wetlands in southern portion of Florida.

Secondly, the marbled crayfish clonal genome was a result of an evolutionary event representing a unique speciation process, which was suggested to occur in *P. fallax* genome. This event gave rise to advantageous asexual lineage, which under inspection based on anthropogenic activity was found to be spread out worldwide, forming an independent habitat which was separated into evolutionary young species named *P. virginalis* (Lyko, 2017). The observation that marbled crayfish is a direct descendant of *P. fallax* contradicts the initial hypothesis for a hybrid-like speciation event and raised the necessity of in-depth genomic analysis of *P. fallax* Everglades populations. Thus, heterozygosity and allele frequency distributions of *P. fallax* specimens were analysed which resulted in detection of triploid genomes among Everglades populations. These triploid specimens were the same specimens in which the parental haplotypes of marbled crayfish were identified. Distribution of biallelic and triallelic SNPs in *P. virginalis* was similar to the one found in triploid *P. fallax* samples. SNPs were mainly biallelic, demonstrating high similarities in the structure of genomes. For the other *P. fallax* specimens sampled from diverse populations, genome heterozygosity estimation resulted in detection of distinct heterozygosity levels. Highest level of heterozygosity was found in triploid animals from Everglades population. These triploid individuals were highly similar, but not identical to *P. virginalis*, due to single nucleotide variations in the mitochondrial and nuclear genomes. The evidence for existence of marbled crayfish population in the native *P. fallax* habitat was not found, and according to recently published data on sex ratio within *P. fallax* populations, the female ratio did not differ from the expected (55.3%). The observed sex ratio was similar to other freshwater crayfish species from Florida (Vogt, 2019), however, unpublished data provided by our collaboration partner (Dr. Nathan J. Dorn from Florida Atlantic University) brings the evidence for female-biased sex ratio of 72.2% at one of the locations with a triploid specimen. The different observation has been made in another location of triploid specimen during the collection period over two years starting in November, 2007. It was characterized by expected (51%) female ratio (van der Heiden and Dorn, 2017).

According to available information and results of comparative analysis of marbled crayfish and its parental species *P. fallax* the conclusion has been made for possible appearance of parthenogenesis, which might arise in triploid *P. fallax* populations. The rare detection of triploid animals among *P. fallax* populations could be explained by elimination of parthenogens by environmental or competitive factors. These observations are consistent with the "Tangled Bank" hypothesis, which emphasizes that the resource competition can favor sex, where the offspring of sexually reproducing species/lines have genetically diverse genotypes and greater chance of survival in long term (Bell, 1982).

Hereby, marbled crayfish originates from triploid *P. fallax* and the current heterogeneous populations of *P. fallax* can be a possible source of other asexual lineages, but this inference requires further confirmation. The way of reproduction of found triploid *P. fallax* individuals from Everglades population remains to be investigated. The differences that exist between the triploid *P. fallax* and *P. virginalis* might soon be answered in detail and help to understand of marbled crayfish speciation.

4.4 Analysis of *P. fallax* population structure

The new genome reference was used for the population structure analysis of *P. fallax*, which resulted in defining four major populations within Florida. The identified genetically distinct populations were broadly aligned with the four major water catchment areas in Florida: Suwannee River, St. John's River, Southwest and Everglades. Interestingly, among all analyzed areas, specimens from three locations in Florida showed a local sympatry of the Southwest and Everglades populations, which explains an appearance of various heterozygosity within populations.

4.5 Parthenogenesis and clonal evolution

As a recent parthenogen and due to its following properties, marbled crayfish became a popular subject of studies in laboratories. First, the high genetic identity observed within marbled crayfish generations indicates the presence of apomictic mode of parthenogenic reproduction (Gutekunst et al., 2018). Microsatellite markers covered heterozygous loci in the analysis of various generations revealed the identical sequences bringing the conclusion, that genetic information was not recombined. Second, the absence of meiosis, which was inferred during oocyte maturation based on the histological studies of nuclei (Vogt et al., 2004). Lately, the histological study with the application of immunohistochemistry brought the first evidence of the presence of meiosis-like incomplete division. In that study, the destabilization of chromosome behaviour during parthenogenic oogenesis was described and referred to gonomery, which is commonly known for apomictic parthenogenic polyploid species. (Kato et al., 2016).

Similar to marbled crayfish, with the availability of complete genome sequences of other parthenogenetic animals, the unique features have become known, such as the presence of allelic regions on the same chromosome in *A. vaga* (Schwander et al., 2011) or substantial heterozygosity, combined with the loss of key meiosis genes in *D. pachys* and *D. coronatus* (Birky, 1996; Welch and Meselson, 2000). However, the availability of genome sequences and their structure raise the relevant questions regarding involved mechanisms of genome reproduction. The study of parthenogenesis opened discussion about genetic diversity in asexual species, their unique evolution and mechanisms of adaptation for species survival. An assumption, where at the absence of meiosis, parthenogens are expected to accumulate allele-independent mutations and increase their genome-wide allele sequence divergence (ASD) known as Meselson effect. In this concept, the ASD is expected to be higher than the divergence in between individuals (Birky, 1996). However, the evidence an increase of inter-allelic divergence was not found in ancient ameiotic asexuals, such as bdelloid rotifer *A. vaga* and the observed gene conversion was suggested to be responsible for limitation of deleterious mutations accumulation in their genome (Flot et al., 2013). Gene conversion can be explained as a result of ameiotic recombination, which leads to allele homogenization or replacing to another allelic version, avoiding Muller's ratchet (Flot et al., 2013). Another process which leads to homogenization is known as Loss Of Heterozygosity (LOH) which can be an important force for inducing variation and may contribute to the limited longevity of asexual species (Forche et al., 2011; Flynn et al., 2017). However, some other asexual species, like *Apis mellifera capensis* and *Daphnia magna* control LOH by reducing of recombination rate for maintaining of high level of heterozygosity (Dukić et al., 2019; Baudry et al., 2004). Similarly to other organisms, LOH was not detected in the marbled crayfish genome, demonstrating resembled genome-wide variant allele frequency for individuals from diverse generations.

Opposite to genetically homogeneous *Daphnia magna* individuals, which can be explained by short observation time between generations, analysis of LOH in marbled crayfish genome included the genome of oldest specimen, demonstrating the most distant generation towards currently existing European populations (Dukić et al., 2019).

Observed results are consistent with apomictic parthenogenesis which was previously suggested in marbled crayfish. Therefore, this allows to infer that genetic variation in marbled crayfish is driven by natural mutational activity. Offspring of independent populations are assumed to gain a lineage-specific mutational pattern in their genome, which was shown in this PhD project. Thus, genome-wide analysis of SNVs confirmed a specifically close genetic relationship within a single population (Figure 13) and emerging differentiation of the marbled crayfish genome in geographically separated populations (Figure 15).

To summarize, this PhD project provides a genetic analysis of the marbled crayfish with a monoclonal population structure. It can be concluded, that a low number of population-specific mutations is consistent with the young age of the European and Malagasy populations (Jones et al., 2009; Kawai et al., 2009; Chucholl et al., 2010).

4.6 Rapid growth of marbled crayfish populations

Parthenogens are often characterized by rapid population growth due to evolutionary successful genotype (via mechanisms discussed in previous section) and high fecundity. Niche and adaptation range are the key ecological factors in population dynamics and intra-/interspecific competition. High fecundity is the result of asexual reproduction, where population consists only all-female individuals and saves the energy on male production. The concept of two-fold “cost of producing males” was discussed in several studies dedicated to comparison of reproduction strategies. Surprisingly, it is a potent factor favoring the clonal reproduction. 50% of individuals in sexual population do not contribute to the fecundity of this group and thus reduce overall fecundity of population.

All-female marbled crayfish populations survive in a broad range of ecological environments and demonstrate a high population density, which was recently observed in many populations in Madagascar (Andriantsoa et al., 2019; Gutekunst et al., 2018) and in Reilingen Lake, analyzed in this PhD project. The estimation of marbled crayfish population size in Lake Reilingen is essential for the assessment of intrinsic population growth rate. Based on known biophysiological features (Vogt et al., 2004) and detected genomic variability of marbled crayfish, the current population dynamics for Reilingen population apparently follows the exponential growth model. The exponential growth is density-independent and exists only during relevant period of the population dynamics (Ciros-Pérez et al., 2001). Thereby, current marbled crayfish population in Reilingen demonstrates a rapid growth, which probably represents an initial step of population dynamics. However, under habitat resource limitations the population is expected to reach saturation and according to the intraspecific competition, a certain constant of population size should be established. Thus, intraspecific competition under high density prevents clones from realization of their two-fold advantage and may result in establishing a drastically new population structure (Doncaster et al., 2000).

4.7 Other applications of the refined marbled crayfish genome assembly in research

The clonal genome of marbled crayfish particularly makes it a promising model organism for epigenetics research. In spite of genetic homogeneity, marbled crayfish have shown a substantial degree of phenotypic variation which are suggested to be controlled via epigenetic regulation.

In the previous studies, many of epigenetic analysis were limited by quality and annotation of the fragmented draft genome assembly. Recently, the comparative analysis of marbled crayfish and *P. fallax* genome methylation was performed, where the refined version of reference genome (v.1) was used for mapping. Particularly, the methylation patterns in triploid *P. fallax* individuals were in the scope of interest. Among analyzed *P. fallax* samples, only triploid individuals demonstrated the hypomethylated DNA across all genomic features, which resembles the pattern of hypomethylation of the marbled crayfish genome (Gatzmann et al., 2018). These findings demonstrate the appearance of large-scale epigenetic changes in the triploid *P. fallax* genome which can be a notable feature in the evolution of the nascent *P. virginalis* genome.

Another important direction in research of adaptability and invasiveness is the functional analysis of epigenetic mechanisms, which occurs in the marbled crayfish model organism. Thus, the inactivation of various epigenetic modifier genes, such as DNA methyltransferases, DNA demethylases, and histone modifying enzymes can provide an important information about their functionality (Carneiro and Lyko, 2020). Thereby, a high quality genome annotation is a key need for epigenetic research of marbled crayfish in order to understand its adaptation and variability.

In summary, the new reference genome of marbled crayfish is the foundation for study of remarkable adaptability, phenotypic plasticity and invasive capability driven by epigenetic mechanisms in the unique monoclonal species.

4.8 Outlook

A finished, accurate reference genome is essential for genome-wide analyses, functional annotation of genome and contribution to genome evolution studies. Particularly, complete marbled crayfish reference genome provides some evidences to the structure of incomplete genomes of other crustaceans. Substantial progress has been made in understanding of marbled crayfish genome evolution, origination and genomic variations within clonal lineages owing to available refined version of the genome.

4.8.1 Marbled crayfish genome assembly v.1

The following improvement of genome assembly includes the genome phasing, which represents a big challenge in polyploid genome assembly. Previously established SNP-based phasing approaches are useful for phasing via comparisons of short variations (SNPs/indels) among two haplotypes in diploids, but they have a limited power to identify the major structural variations within a polyploid genome (Yang et al., 2017). The application of existing Hi-C scaffolding programs for polyploid genomes results in production of chimeric chromosomes, which have joined allelic contigs from different haplotypes. A recent improved scaffolding tool for autopolyploids is based on allele-aware algorithms using Hi-C paired-end reads, which produces the chromosomal-scale scaffolds (Zhang et al., 2019). It has been applied on an autotetraploid and an autooctoploid sugar-cane genome and successfully constructed the phased chromosomal-level assemblies, demonstrating a powerful tool for overcoming obstacles in assembling complex polyploid genomes and can be considered for for marbled crayfish genome scaffolding in the future. Particular importance of allele-aware scaffolding is improvement of mosaic assembly of the marbled crayfish genome for separation of individual AA'B haplotypes.

Another important refinement is related to genome annotation. The critical step in genome annotation is the manual curation, which assesses and improves the gene prediction accuracy. A better quality of the marbled crayfish genome annotation will significantly enhance the value of gene predictions and will be beneficial for the research community. The current results of automated annotation require further corrections and are publicly available for manual curation.

4.8.2 Marbled crayfish as a promising model organism

After release of the refined genome, another important questions concerning the unique features of marbled crayfish genome are expected to be raised in the future. For instance, the specific mechanisms which distinguish marbled crayfish from triploid *P. fallax* individuals would be the main scope of interest. Particularly valuable will be a comprehensive study of the functional relevance of repetitive DNA in the genome. For instance, activity of transposable elements (TE) was shown to be connected with parthenogenesis and also suggested to be involved in imposing a mutational burden in recent asexuals (Sullender and Crease, 2001; Schaack et al., 2010; Nuzhdin and Petrov, 2003; Kozlowski et al., 2020). Moreover, in very recent study TE were found to form the insertions within genic and in the upstream regulatory regions. This was presumed to have a functional impact and play a role in the genome plasticity (Kozlowski et al., 2020). Apart from TE, analysis of copy-number variations (CNV), large-scale structural variations (SVs) could be involved in the remarkable adaptability of marbled crayfish likewise it was suggested in the study of another clonal species *M. incognita* (Koutsovoulos et al., 2020).

The genome analysis of the sexually reproducing *P. fallax* species and search for distinguishing genomic features, like meiosis-related genes can provide some insights on the mechanisms related to obligate apomictic parthenogenesis.

Finally, the cost of asexual marbled crayfish reproduction has to be investigated. Other aspects of genomic, epigenetic and metabolic variabilities and their connection with biological traits such as chromosomal behaviour during oogenesis, geographical distribution and subsequent adaptations should be investigated in the future research. Elucidation the role of ecology for sexual and asexual reproduction might be useful in search of answers for the evolution of sex in the organism's genomes.

5 Additional contributions

The additional contribution has been made in a development of computational pipeline for direct detection of RNA modifications by nanopore sequencing application. Oxford Nanopore Technology (ONT) promoted the extensive development of nanopore sequencing, which apart from generation of long reads in real time mode has other advantages. First of all, it enables the easy-setup of sequencing machinery, which does not require the capital cost for installation and its compact size makes ONT suitable for every lab. Secondly, nanopore sequencing allows the direct detection of nucleotide modifications. These remarkable advantages made ONT useful in application for the direct detection of tRNA modifications in the research group of RNA modification at DKFZ, Epigenetics division. The first direct detection of RNA modification was performed on rRNA molecule (Smith et al., 2019) and on mRNA (Garalde et al., 2018), but has not been applied for direct detection of tRNA modifications. The project on tRNA modifications was guided by Dr. Francesca Tuorto in which the major focus was the analysis of queuosine modification and its role in regulation of protein translation.

The bioinformatic side of the project included the establishment of the pipeline for base calling and detection of signals specific for queuosine modifications within designed oligo sequence. The strategy of queuosine modification detection was based on signal perturbation in the pairwise comparison of currents profiles of modified and non-modified sequences. In the modified molecule the current profile showed the signal perturbation at the position 33, where the nucleotide modified by queuosine molecule. However, the read coverage in unmodified molecule sequencing was not sufficient for the convincing inference. Apart from sequencing of designed oligos with queuosine modification, the other native tRNA modifications were aimed to be detected by direct nanopore sequencing. The pool of native diverse tRNAs were prepared for the single library run. The high number of natural modifications of tRNA molecules and high error rate of base calling in nanopore sequencing resulted in generation of noisy signals and thus, the base calling of low efficiency. However, multiple runs of sequencing and less stringent filtering criteria enabled the recognition of majority of tRNA types. The number of modified nucleotides of the tRNA molecule was high which causes the miscalling of bases. Thus, the direct nanopore sequencing and modification detection require benchmarking for each type of non-canonical modification which have not been included in the pipeline of RNA modifications detection.

5.1 List of publications containing personal contributions

- Olena Maiakovska, Ranja Andriantsoa, Sina Tonges, Carine Legrand, Julian Gutekunst, Katharina Hanna, Lucian Parvulescu, Roman Novitsky, Andras Weiperth, Arnold Sciberras, Alan Deidun, Fabio Ercoli, Antonin Kouba, Frank Lyko, Genomic variation in the monoclonal marbled crayfish (Procambarus virginalis) Communications Biology, 2020 [accepted]

- Julian Gutekunst, Olena Maiakovska, Geetha Venkatesh, Katharina Hanna, Hannes Horn, Stephan Wolf, Christopher E. Skelton, Nathan J. Dorn, Frank Lyko, Nascent genome evolution and speciation of the parthenogenetic marbled crayfish [in preparation]

A Appendix

Table 10: **PacBio sequencing results.** Statistics of generated CCS and subreads from the PacBio SEQUEL platform

ILSe ID	ASLR	CCS Reads	CCS Bases	Sub Reads
9976	AS-213883-LR-31767	79,893	399,493,263	3,267,709,171
9976	AS-213883-LR-31768	109,151	507,396,960	4,083,444,122
9977	AS-213888-LR-31769	157,489	905,026,471	6,197,693,915
9977	AS-213888-LR-31770	166,891	855,338,399	5,787,145,424
9977	AS-213888-LR-31771	164,421	846,014,400	5,654,783,289
9956	AS-214589-LR-31818	123,704	511,102,102	3,868,786,366
9955	AS-214587-LR-31819	95,210	348,452,776	2,986,569,954
8630	AS-187423-LR-30941	96,225	556,632,697	3,771,631,765
10670	AS-229076-LR-34711	227,493	799,192,672	8,379,050,181
10670	AS-229078-LR-34712	138,104	528,008,192	3,327,489,819
10670	AS-229080-LR-34713	179,626	644,163,919	3,985,650,209
10670	AS-229082-LR-34714	140,840	486,941,417	3,481,517,359
10670	AS-229084-LR-34715	260,344	929,101,020	12,037,917,866
10670	AS-229086-LR-34716	208,864	756,228,956	7,407,710,158
10670	AS-229088-LR-34717	210,110	793,754,037	8,234,062,423
10670	AS-229090-LR-34718	198,353	720,623,585	7,446,019,680
10892	AS-245833-LR-35564	147,318	728,740,808	6,347,226,199
10892	AS-245833-LR-35565	152,030	742,246,983	6,392,169,533
10892	AS-245833-LR-35566	160,195	891,183,518	7,237,732,493
10892	AS-245833-LR-35567	156,999	876,109,568	7,081,458,212
10892	AS-245833-LR-35568	166,731	934,741,359	7,134,259,292
10892	AS-245833-LR-35569	189,505	1,054,766,389	7,901,778,324
10892	AS-245833-LR-35570	193,068	1,088,839,703	7,712,804,715
10892	AS-245833-LR-35571	202,534	1,114,042,469	8,370,188,770
10892	AS-245833-LR-35659	156,427	674,555,075	6,198,609,092
10892	AS-245833-LR-35660	199,794	866,227,497	7,758,093,218
10892	AS-250879-LR-34794	183,559	775,976,627	7,084,473,077
10891	AS-250879-LR-35661	258,231	1,000,394,225	8,518,727,776
10891	AS-250879-LR-35662	258,231	1,000,394,225	8,047,402,192
10891	AS-250879-LR-35756	263,516	1,003,413,139	9,795,595,345
10891	AS-250879-LR-35791	162,708	691,287,839	6,431,340,980
10891	AS-250879-LR-35792	196,608	822,331,891	7,590,043,957
10891	AS-250879-LR-35793	224,281	908,229,563	8,491,235,779
10891	AS-250879-LR-35795	166,624	705,592,013	6,434,877,485
10891	AS-250879-LR-35796	191,813	812,806,747	7,633,386,938
10891	AS-250879-LR-35797	227,651	927,461,784	8,256,875,999
10655	AS-228802-LR-34720	287,888	963,926,623	7,704,304,275
Sum		6,702,429	29,170,738,911	248,039,765,352

List of Figures

List of Tables

References

Shanika L Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, 21(1):1–16, 2020.

Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

Ranja Andriantsoa, Sina Tönges, Jörn Panteleit, Kathrin Theissinger, Vitor Coutinho Carneiro, Jeanne Rasamy, and Frank Lyko. Ecological plasticity and commercial impact of invasive marbled crayfish populations in madagascar. *BMC ecology*, 19(1):1–10, 2019.

Kin Fai Au, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. Improving pacbio long read accuracy by short read alignment. *PloS one*, 7(10):e46679, 2012.

Marina Barba, Henryk Czosnek, and Ahmed Hadidi. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses*, 6(1):106–136, 2014.

Emmanuelle Baudry, Per Kryger, Mike Allsopp, Nikolaus Koeniger, Dominique Vautrin, Florence Mougel, Jean-Marie Cornuet, and Michel Solignac. Whole-genome scan in thelytokous-laying workers of the cape honeybee (apis mellifera capensis): central fusion, reduced recombination rates and centromere mapping using half-tetrad analysis. *Genetics*, 167(1):243–252, 2004.

Graham Bell. *The Masterpiece of Nature:: The Evolution and Genetics of Sexuality*. CUP Archive, 1982.

David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59, 2008.

C William Birky. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics*, 144(1):427–437, 1996.

Marten Boetzer, Christiaan V Henkel, Hans J Jansen, Derek Butler, and Walter Pirovano. Scaffolding pre-assembled contigs using sspace. *Bioinformatics*, 27(4):578–579, 2011.

Patrik Bohman, Lennart Edsman, Peer Martin, and Gerhard Scholtz. The first marmorkrebs (decapoda: Astacida: Cambaridae) in scandinavia. 2013.

Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

Jacobus J Boomsma, Sean G Brady, Robert R Dunn, Juergen Gadau, Juergen Heinze, Laurent Keller, Corrie S Moreau, Nathan J Sanders, Lukas Schrader, Ted R Schultz, et al. The global ant genomics alliance (gaga), 2017.

Vitor Coutinho Carneiro and Frank Lyko. Rapid epigenetic adaptation in animals and its role in invasiveness. *Integrative and Comparative Biology*, 2020.

Mark Chaisson, Pavel Pevzner, and Haixu Tang. Fragment assembly with short reads. *Bioinformatics*, 20(13):2067–2074, 2004.

Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611, 2015a.

Mark JP Chaisson, Richard K Wilson, and Evan E Eichler. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11):627–640, 2015b.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

Erwin Chargaff, Rakoma Lipshitz, Charlotte Green, and ME Hodes. The composition of the desoxyribonucleic acid of salmon sperm. *Journal of Biological Chemistry*, 192(1): 223–230, 1951.

Ying Chen, Fan Nie, Shang-Qian Xie, Ying-Feng Zheng, Thomas Bray, Qi Dai, Yao-Xin Wang, Jian-feng Xing, Zhi-Jian Huang, De-Peng Wang, et al. Fast and accurate assembly of nanopore reads via progressive error correction and adaptive read selection. *bioRxiv*, 2020.

Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, et al. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, 10(6):563–569, 2013.

Christoph Chucholl. Marbled crayfish gaining ground in europe: the role of the pet trade as invasion pathway. *Freshwater crayfish: a global overview*, pages 83–114, 2015.

Christoph Chucholl, Michael Pfeiffer, et al. 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, 2010.

Jorge Ciros-Pérez, María José Carmona, and Manuel Serra. Resource competition between sympatric sibling rotifer species. *Limnology and Oceanography*, 46(6):1511–1523, 2001.

John K Colbourne, Michael E Pfrender, Donald Gilbert, W Kelley Thomas, Abraham Tucker, Todd H Oakley, Shinichi Tokishita, Andrea Aerts, Georg J Arnold, Malay Kumar Basu, et al. The ecoresponsive genome of daphnia pulex. *Science*, 331(6017): 555–561, 2011.

Luca Comai. The advantages and disadvantages of being polyploid. *Nature reviews genetics*, 6(11):836–846, 2005.

1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061, 2010.

Marija Cvitanić, Sandra Hudina, and Ivana Maguire. Reproductive cycle of the marble crayfish from an established population in croatia. In *21st Symposium of the International Association of Astacology-Program and book of abstracts*, page 48, 2016.

Rami A Dalloul, Julie A Long, Aleksey V Zimin, Luqman Aslam, Kathryn Beal, Le Ann Blomberg, Pascal Bouffard, David W Burt, Oswald Crasta, Richard PMA Crooijmans, et al. Multi-platform next-generation sequencing of the domestic turkey (meleagris gallopavo): genome assembly and analysis. *PLoS Biol*, 8(9):e1000475, 2010.

Alan Deidun, Arnold Sciberras, Justin Formosa, Bruno Zava, Gianni Insacco, Maria Corsini-Foka, and Keith A Crandall. Invasion by non-indigenous freshwater decapods of malta and sicily, central mediterranean sea. *Journal of Crustacean Biology*, 38(6): 748–753, 2018.

Gennady Denisov, Brian Walenz, Aaron L Halpern, Jason Miller, Nelson Axelrod, Samuel Levy, and Granger Sutton. Consensus generation and variant detection by celera assembler. *Bioinformatics*, 24(8):1035–1040, 2008.

Stéphane Deschamps and Matthew A Campbell. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular breeding*, 25(4): 553–570, 2010.

C Patrick Doncaster, Graeme E Pound, and Simon J Cox. The ecological cost of sex. *Nature*, 404(6775):281–285, 2000.

Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, et al. De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, 356(6333):92–95, 2017.

Marinela Dukić, Daniel Berner, Christoph R Haag, and Dieter Ebert. How clonal are clones? a quest for loss of heterozygosity during asexual reproduction in daphnia magna. *Journal of evolutionary biology*, 32(6):619–628, 2019.

John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

Robert J Elshire, Jeffrey C Glaubitz, Qi Sun, Jesse A Poland, Ken Kawamoto, Edward S Buckler, and Sharon E Mitchell. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PloS one*, 6(5):e19379, 2011.

Christine G Elsik, Kim C Worley, Anna K Bennett, Martin Beye, Francisco Camara, Christopher P Childers, Dirk C de Graaf, Griet Debyser, Jixin Deng, Bart Devreese, et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC genomics*, 15(1):86, 2014.

Genomics England. The 100,000 genomes project. *The*, 100:0–2, 2016.

Adam C English, Stephen Richards, Yi Han, Min Wang, Vanesa Vee, Jiaxin Qu, Xiang Qin, Donna M Muzny, Jeffrey G Reid, Kim C Worley, et al. Mind the gap: upgrading genomes with pacific biosciences rs long-read sequencing technology. *PloS one*, 7(11): e47768, 2012.

Fabio Ercoli, Katrin Kaldre, Tiit Paaver, and Riho Gross. First record of an established marbled crayfish procambarus virginalis (lyko, 2017) population in estonia. *BioInvasions Records*, 8(3), 2019.

Cassandra Falckenhayn. *The methylome of the marbled crayfish Procambarus virginalis*. PhD thesis, 2017.

Varvara Fazalova and Bruno Nevado. Low spontaneous mutation rate and pleistocene radiation of pea aphids. *Molecular biology and evolution*, 37(7):2045–2051, 2020.

Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

Jean-François Flot, Boris Hespeels, Xiang Li, Benjamin Noel, Irina Arkhipova, Etienne GJ Danchin, Andreas Hejnol, Bernard Henrissat, Romain Koszul, Jean-Marc Aury, et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer adineta vaga. *Nature*, 500(7463):453–457, 2013.

Jullien M Flynn, Frederic JJ Chain, Daniel J Schoen, and Melania E Cristescu. Spontaneous mutation accumulation in daphnia pulex in selection-free vs. competitive environments. *Molecular Biology and Evolution*, 34(1):160–173, 2017.

A Forche, D Abbey, T Pisithkul, MA Weinzierl, T Ringstrom, D Bruck, K Petersen, and J Berman. Stress alters rates and types of loss of heterozygosity in candida albicans. *MBio*, 2(4), 2011.

Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct rna sequencing on an array of nanopores. *Nature methods*, 15(3):201, 2018.

E Garrison and G Marth. Freebayes. arxiv preprint1207. 3907 [q-bio. gn][internet], 2012.

Fanny Gatzmann, Cassandra Falckenhayn, Julian Gutekunst, Katharina Hanna, Günter Raddatz, Vitor Coutinho Carneiro, and Frank Lyko. The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & chromatin*, 11(1):57, 2018.

Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal*, 18:9–19, 2020.

Frances Goudie, Michael H Allsopp, Madeleine Beekman, Peter R Oxley, Julianne Lim, and Benjamin P Oldroyd. Maintenance and loss of heterozygosity in a thelytokous lineage of honey bees (apis mellifera capensis). *Evolution: International Journal of Organic Evolution*, 66(6):1897–1906, 2012.

Julian Gutekunst, Ranja Andriantsoa, Cassandra Falckenhayn, Katharina Hanna, Wolfgang Stein, Jeanne Rasamy, and Frank Lyko. Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nature ecology & evolution*, 2(3):567, 2018.

i5K Consortium. The i5k initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, 104(5):595–600, 2013.

Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338–345, 2018.

Kaitlyn E Johnson, Grant Howard, William Mo, Michael K Strasser, Ernesto ABF Lima, Sui Huang, and Amy Brock. Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an allee effect. *PLoS biology*, 17(8): e3000399, 2019.

Julia PG Jones, Jeanne R Rasamy, Andrew Harvey, Alicia Toon, Birgit Oidtmann, Michele H Randrianarison, Noromalala Raminosoa, and Olga R Ravoahangimalala.

The perfect invader: a parthenogenic crayfish poses a new threat to madagascar's freshwater biodiversity. *Biological Invasions*, 11(6):1475–1482, 2009.

Young Seok Ju, Jong-Il Kim, Sheehyun Kim, Dongwan Hong, Hansoo Park, Jong-Yeon Shin, Seungbok Lee, Won-Chul Lee, Sujung Kim, Saet-Byeol Yu, et al. Extensive genomic and transcriptional diversity identified through massively parallel dna and rna sequencing of eighteen korean individuals. *Nature genetics*, 43(8):745–752, 2011.

Miku Kato, Chizue Hiruta, and Shin Tochinai. The behavior of chromosomes during parthenogenetic oogenesis in marmorkrebs procambarus fallax f. virginalis. *Zoological science*, 33(4):426–430, 2016.

Tadashi Kawai, Gerhard Scholtz, Shinsuke Morioka, Fihaonantsoa Ramanamandimby, Chris Lukhaup, and Yukio Hanamura. Parthenogenetic alien crayfish (decapoda: Cambaridae) spreading in madagascar. *Journal of Crustacean Biology*, 29(4):562–567, 2009.

Klaus-Peter Koepfli, Benedict Paten, Genome 10K Community of Scientists, and Stephen J O'Brien. The genome 10k project: a way forward. *Annu. Rev. Anim. Biosci.*, 3(1):57–111, 2015.

Sergey Koren, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, 30(7):693–700, 2012.

Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, 27(5):722–736, 2017.

Georgios D Koutsovoulos, Eder Marques, Marie-Jeanne Arguel, Laurent Duret, Andressa CZ Machado, Regina MDG Carneiro, Djampa K Kozlowski, Marc Bailly-Bechet, Philippe Castagnone-Sereno, Erika VS Albuquerque, et al. Population genomics supports clonal reproduction and multiple independent gains and losses of parasitic abilities in the most devastating nematode pest. *Evolutionary applications*, 13(2):442–457, 2020.

Djampa KL Kozlowski, Rahim Hassanaly-Goulamhoussen, Martine Da Rocha, Georgios D Koutsovoulos, Marc Bailly-Bechet, and Etienne GJ Danchin. Transposable elements are an evolutionary force shaping genomic plasticity in the parthenogenetic root-knot nematode meloidogyne incognita. *bioRxiv*, 2020.

CJ Krebs. Estimating abundance in animal and plant populations. *Ecological Methodology,*, pages 24–77, 2016.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

Thomas Laver, J Harrison, PA O'neill, Karen Moore, Audrey Farbos, Konrad Paszkiewicz, and David J Studholme. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection and quantification*, 3:1–8, 2015.

Hayan Lee, James Gurtowski, Shinjae Yoo, Maria Nattestad, Shoshana Marcus, Sara Goodwin, W Richard McCombie, and Michael C Schatz. Third-generation sequencing and the future of genomics. *BioRxiv*, page 048603, 2016.

Tom Levy, Ohad Rosen, Ohad Simons, Amit Savaya Alkalay, and Amir Sagi. The gene encoding the insulin-like androgenic gland hormone in an all-female parthenogenetic crayfish. *PloS one*, 12(12):e0189982, 2017.

Harris A Lewin, Gene E Robinson, W John Kress, William J Baker, Jonathan Coddington, Keith A Crandall, Richard Durbin, Scott V Edwards, Félix Forest, M Thomas P Gilbert, et al. Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, 2018.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Ruiqiang Li, Wei Fan, Geng Tian, Hongmei Zhu, Lin He, Jing Cai, Quanfei Huang, Qingle Cai, Bo Li, Yinqi Bai, et al. The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):311–317, 2010.

Xingyu Liao, Min Li, You Zou, Fang-Xiang Wu, Jianxin Wang, et al. Current challenges and solutions of de novo assembly. *Quantitative Biology*, pages 1–20, 2019.

B Lipták, A Mrugała, L Pekárik, A Mutkovič, and D Grul'a. Petrusek a, kouba a (2016) expansion of the marbled crayfish in slovakia: beginning of an invasion in the danube catchment? journal of limnology 75: 305-312. *Agata Mrugała*, 75(2):305–312, 2016.

Haoxuan Liu, Yanxiao Jia, Xiaoguang Sun, Dacheng Tian, Laurence D Hurst, and Sihai Yang. Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. *Molecular biology and evolution*, 34(1):119–130, 2017.

Frank Lyko. The marbled crayfish (decapoda: Cambaridae) represents an independent new species. *Zootaxa*, 4363(4):544–552, 2017.

Michael Lynch. Destabilizing hybridization, general-purpose genotypes and geographic parthenogenesis. *The Quarterly Review of Biology*, 59(3):257–290, 1984.

Tiffani M Manteuffel-Ross, Eric Stolen, and C Ross Hinkle. Abundance and habitat associations of two florida crayfishes, procambarus paeninsulanus (faxon, 1914) and p. fallax (hagen, 1870)(decapoda: Astacoidea), assessed with n-mixture modeling. *Journal of Crustacean Biology*, 38(3):285–294, 2018.

Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

Peer Martin, Nathan J Dorn, Tadashi Kawai, Craig van der Heiden, and Gerhard Scholtz. The enigmatic marmorkrebs (marbled crayfish) is the parthenogenetic form of procambarus fallax (hagen, 1870). *Contributions to Zoology*, 79(3):107–118, 2010.

Peer Martin, Sven Thonagel, and Gerhard Scholtz. The parthenogenetic m armorkrebs (m alacostraca: D ecapoda: C ambaridae) is a triploid organism. *Journal of Zoological Systematics and Evolutionary Research*, 54(1):13–21, 2016.

Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303, 2010.

André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome biology*, 12(11):R112, 2011.

Ursula Mittwoch. Parthenogenesis. *Journal of Medical Genetics*, 15(3):165, 1978.

Joaquín Muñoz, Anurag Chaturvedi, Luc De Meester, and Lawrence J Weider. Characterization of genome-wide snps for the water flea daphnia pulicaria generated by genotyping-by-sequencing (gbs). *Scientific reports*, 6:28569, 2016.

Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl_2):ii79–ii85, 2005.

Eugene W Myers Jr. A history of dna sequence assembly. *It-Information Technology*, 58 (3):126–132, 2016.

Benjamin B Normark. The evolution of alternative genetic systems in insects. *Annual Review of Entomology*, 48(1):397–423, 2003.

Roman A Novitsky and Mikhail O Son. The first records of marmorkrebs [procambarus fallax (hagen, 1870) f. virginalis](crustacea, decapoda, cambaridae) in ukraine. *Ecologica Montenegrina*, 5:44–46, 2016.

Sergey V Nuzhdin and Dmitri A Petrov. Transposable elements in clonal lineages: lethal hangover from sex. *Biological Journal of the Linnean Society*, 79(1):33–41, 2003.

Lucian Pârvulescu, Andrei Togor, Sandra-Florina Lele, Sebastian Scheu, Daniel Șinca, and Jörn Panteleit. First established population of marbled crayfish procambarus fallax (hagen, 1870) f. virginalis (decapoda, cambaridae) in romania. *BioInvasions Record*, 6 (4), 2017.

Jiří Patoka, Miloš Buřič, Vojtěch Kolář, Martin Bláha, Miloslav Petrtỳl, Pavel Franta, Robert Tropek, Lukáš Kalous, Adam Petrusek, and Antonín Kouba. Predictions of marbled crayfish establishment in conurbations fulfilled: Evidences from the czech republic. *Biologia*, 71(12):1380–1385, 2016.

Valentina Peona, Matthias H Weissensteiner, and Alexander Suh. How complete are "complete" genome assemblies?—an avian perspective. *Molecular ecology resources*, 18(6): 1188–1195, 2018.

Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An eulerian path approach to dna fragment assembly. *Proceedings of the national academy of sciences*, 98(17):9748–9753, 2001.

Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, 2009.

Liam J Revell. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, 3(2):217–223, 2012.

Ben M Sadd, Seth M Barribeau, Guy Bloch, Dirk C De Graaf, Peter Dearden, Christine G Elsik, Jürgen Gadau, Cornelis JP Grimmelikhuijzen, Martin Hasselmann, Jeffrey D Lozier, et al. The genomes of two key bumblebee species with primitive eusocial organization. *Genome biology*, 16(1):1–32, 2015.

Leena Salmela and Eric Rivals. Lordec: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

Frederick Sanger, Gilian M Air, Bart G Barrell, Nigel L Brown, Alan R Coulson, J_ C_ Fiddes, CA Hutchison, Patrick M Slocombe, and Mo Smith. Nucleotide sequence of bacteriophage φx174 dna. *nature*, 265(5596):687–695, 1977a.

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

Sarah Schaack, Ellen J Pritham, Abby Wolf, and Michael Lynch. Dna transposon dynamics in populations of daphnia pulex with and without sex. *Proceedings of the Royal Society B: Biological Sciences*, 277(1692):2381–2387, 2010.

Eric E Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human molecular genetics*, 19(R2):R227–R240, 2010.

Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173, 2010.

Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, et al. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5):849–864, 2017.

Gerhard Scholtz, Anke Braband, Laura Tolley, André Reimann, Beate Mittmann, Chris Lukhaup, Frank Steuerwald, and Günter Vogt. Parthenogenesis in an outsider crayfish. *Nature*, 421(6925):806–806, 2003.

Stephan C Schuster, Webb Miller, Aakrosh Ratan, Lynn P Tomsho, Belinda Giardine, Lindsay R Kasson, Robert S Harris, Desiree C Petersen, Fangqing Zhao, Ji Qi, et al. Complete khoisan and bantu genomes from southern africa. *Nature*, 463(7283):943–947, 2010.

Tanja Schwander, Lee Henry, and Bernard J Crespi. Molecular evidence for ancient asexuality in timema stick insects. *Current Biology*, 21(13):1129–1134, 2011.

Robert Seitz, Kathia Vilpoux, Ulrich Hopp, Steffen Harzsch, and Gerhard Maier. Ontogeny of the marmorkrebs (marbled crayfish): a parthenogenetic crayfish with unknown origin and phylogenetic position. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 303(5):393–405, 2005.

Rekha Seshadri, Sinead C Leahy, Graeme T Attwood, Koon Hoong Teh, Suzanne C Lambie, Adrian L Cookson, Emiley A Eloe-Fadrosh, Georgios A Pavlopoulos, Michalis Hadjithomas, Neha J Varghese, et al. Cultivation and sequencing of rumen microbiome members from the hungate1000 collection. *Nature biotechnology*, 36(4):359, 2018.

Jared T Simpson and Mihai Pop. The theory and practice of genome sequence assembly. *Annual review of genomics and human genetics*, 16, 2015.

Andrew M Smith, Miten Jain, Logan Mulroney, Daniel R Garalde, and Mark Akeson. Reading canonical and modified nucleobases in 16s ribosomal rna using nanopore native rna sequencing. *PloS one*, 14(5):e0216709, 2019.

Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

R Stouthamer, JAJ Breeuwer, RF Luck, and JH Werren. Molecular identification of microorganisms associated with parthenogenesis. *Nature*, 361(6407):66–68, 1993.

Barry W Sullender and Teresa J Crease. The behavior of a daphnia pulex transposable element in cyclically and obligately parthenogenetic populations. *Journal of molecular evolution*, 53(1):63–69, 2001.

Mun Hua Tan, Christopher M Austin, Michael P Hammer, Yin Peng Lee, Laurence J Croft, and Han Ming Gan. Finding nemo: hybrid assembly with oxford nanopore and illumina reads greatly improves the clownfish (amphiprion ocellaris) genome assembly. *GigaScience*, 7(3):gix137, 2018.

R Core Team et al. R: A language and environment for statistical computing, 2013.

Picard Tools. By broad institute, 2015.

Maria Tsompana and Michael J Buck. Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, 7(1):1–16, 2014.

Craig A van der Heiden and Nathan J Dorn. Benefits of adjacent habitat patches to the distribution of a crayfish population in a hydro-dynamic wetland landscape. *Aquatic Ecology*, 51(2):219–233, 2017.

Casper J van der Kooi, Cyril Matthey-Doret, and Tanja Schwander. Evolution and comparative ecology of parthenogenesis in haplodiploid arthropods. *Evolution letters*, 1(6): 304–316, 2017.

A Vandel. La parthénogénèse géographique. iv. polyploidie et distribution géographique. *Bull Biol France Belg*, 74:94–l00, 1940.

J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

Günter Vogt. Bimodal annual reproductive pattern in laboratory-reared marbled crayfish. *Invertebrate Reproduction & Development*, 59(4):218–223, 2015.

Günter Vogt. Estimating the young evolutionary age of marbled crayfish from museum samples. *Journal of Natural History*, 53(39-40):2353–2363, 2019.

Günter Vogt, Laura Tolley, and Gerhard Scholtz. Life stages and reproductive components of the marmorkrebs (marbled crayfish), the first parthenogenetic decapod crustacean. *Journal of Morphology*, 261(3):286–311, 2004.

Günter Vogt, Cassandra Falckenhayn, Anne Schrimpf, Katharina Schmid, Katharina Hanna, Jörn Panteleit, Mark Helm, Ralf Schulz, and Frank Lyko. The marbled crayfish as a paradigm for saltational speciation by autopolyploidy and parthenogenesis in animals. *Biology open*, 4(11):1583–1594, 2015.

R Vojkovská, I Horká, E Tricarico, and Z Ďuriš. New record of the parthenogenetic marbled crayfish procambarus fallax f. virginalis from italy. *Crustaceana*, 87(11-12):1386–1392, 2014.

Robert C Vrijenhoek. Factors affecting clonal diversity and coexistence. *American Zoologist*, 19(3):787–797, 1979.

Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

Jun Wang, Wei Wang, Ruiqiang Li, Yingrui Li, Geng Tian, Laurie Goodman, Wei Fan, Junqing Zhang, Jun Li, Juanbin Zhang, et al. The diploid genome sequence of an asian individual. *Nature*, 456(7218):60–65, 2008.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. *Genome research*, 27(5):757–767, 2017.

David B Mark Welch and Matthew Meselson. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science*, 288(5469):1211–1215, 2000.

Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10): 1155–1162, 2019.

David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872–876, 2008.

Jing Qin Wu, Chongmei Dong, Long Song, and Robert F Park. Long-read–based de novo genome assembly and comparative genomics of the wheat leaf rust pathogen puccinia triticina identifies candidates for three avirulence genes. *Frontiers in genetics*, 11:521, 2020.

Enhua Xia, Fangdong Li, Wei Tong, Hua Yang, Songbo Wang, Jian Zhao, Chun Liu, Liping Gao, Yuling Tai, Guangbiao She, et al. The tea plant reference genome and improved gene annotation using long-read and paired-end sequencing data. *Scientific data*, 6(1): 1–9, 2019.

Wei Xue, Jiong-Tang Li, Ya-Ping Zhu, Guang-Yuan Hou, Xiang-Fei Kong, You-Yi Kuang, and Xiao-Wen Sun. L_rna_scaffolder: scaffolding genomes with transcripts. *BMC genomics*, 14(1):604, 2013.

Jun Yang, M-Hossein Moeinzadeh, Heiner Kuhl, Johannes Helmuth, Peng Xiao, Stefan Haas, Guiling Liu, Jianli Zheng, Zhe Sun, Weijuan Fan, et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nature plants*, 3(9):696–703, 2017.

Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

Xingtan Zhang, Shengcheng Zhang, Qian Zhao, Ray Ming, and Haibao Tang. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on hi-c data. *Nature plants*, 5(8):833–845, 2019.

Xingtan Zhang, Ruoxi Wu, Yibin Wang, Jiaxin Yu, and Haibao Tang. Unzipping haplotypes in diploid and polyploid genomes. *Computational and structural biotechnology journal*, 18:66–72, 2020.

Acknowledgements

This thesis has been an exiting and challenging journey and would not have been possible without great support and sincere feedback of many people, to whom I am deeply grateful.

First and foremost, I would like to express the gratefulness to my supervisor Frank Lyko for guiding me through the last three years, for spending an infinite time in correcting my first manuscript, in supporting my progress and who did a great input in my maturation as a scientist.

My special thank to Julian Gutekunst for being a very supportive mentor, a wonderful colleague and for our shared undisguised love for cats.

Thank you to my thesis advisory committee - Benedikt Brors and Christoph Dieterich for your active interest in my research, meaningful feedback, and your guidance.

I feel overwhelmingly grateful to all my dear colleagues and officemates: Vitor for his inspiring and enthusiastic scientific discussions and kindhearted attitude, Geetha for her kindness and support, Kathi for the enormous amount of delicious cakes and her excessive lab work, Sina & Ranja who made a great team of beautiful and talented female scientists, Fanny for being a part of our cheerful team. Thank you all for genuine childishness we had together in teasing and pranking each other, in having memorable cantuccini-coffee breaks. I also wish to express my appreciation to other members of Epigenetics group: Carine for her friendship, support, and encouragement, Günther for his deep scientific questions and helpful discussions, Francesca & Cansu for being the greatest flatmates and the best rowers during our retreat, Sofia & Julian for being my first students to whom I was honoured to provide the bioinformatic support, Manolo for making the most cheerful speeches and all other former and current teammates whom I had chance to meet.

I am greatful to my loving mother and sister for being supportive and encouraging during all these years.

I am thankful to my best friends Anna and Niharika for your unceasing support and making me feel like home wherever I am with you. I am greatful to Sunidhi, Laasya, Vani, Raluca, Ritika, Andrey, Renat, Sid, Alex, Igor and Nastia for being a good friends and sharing the worries on our common path beight PhD students during these years.

To my dear Rangel for a healthy criticism, for his constant support and endless scientific discussions, which inspired me to accomplish my mission to become a scientist.

Thank you all!