The Transcriptomic and Genomic Architecture of Acrididae Grasshoppers

Dissertation

To Fulfil the
Requirements for the Degree of

"Doctor of Philosophy" (PhD)

Submitted to the Council of the Faculty
of Biological Sciences
of the Friedrich Schiller University Jena

by Bachelor of Science, Master of Science, Abhijeet Shah

born on 7th November 1984, Hyderabad, India

Academic reviewers:

1. Prof. Holger Schielzeth, Friedrich Schiller University Jena
2. Prof. Manja Marz, Friedrich Schiller University Jena
3. Prof. Rolf Beutel, Friedrich Schiller University Jena
4. Prof. Frieder Mayer, Museum für Naturkunde Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin
5. Prof. Steve Hoffmann, Leibniz Institute on Aging – Fritz Lipmann Institute, Jena
6. Prof. Aletta Bonn, Friedrich Schiller University Jena

Date of oral defense:

24.02.2020

Table of Contents

Abstract

Genetic polymorphism is described as the variation in DNA sequence between distinct individuals of a given species or population. This polymorphism is reflected from individuals to entire populations, and from single nucleotides to the entire genome spanning billions of base pairs. A fundamental aim of functional genomics is to establish links between genetic polymorphism and phenotypic variation, to explain this observed variation. Recent developments in high throughput sequencing have made it possible to adequately explore this link. My dissertation explores genetic and genomic polymorphism in Gomphocerine grasshoppers, an insect group with unusually large and complex genomes using novel and contemporary transcriptomic and genomic methods.

I have developed a novel method for microsatellite discovery in the club-legged grasshopper (*Gomphocerus sibiricus*) and have created a high-quality reference transcriptome to further facilitate studies. Additionally, I demonstrated the utility of high-throughput sequencing and prior bioinformatic analysis to dramatically further reduce costs and improve efficiency of developing microsatellite markers by reducing primer binding interference. Leveraging the information gathered during *de novo* reference transcriptome development, I demonstrated that the entire mitochondrial genome could also be assembled, and sequence divergence to closely related species and populations estimated. This reference assembly also yielded evidence of the endo-parasite *Wolbachia* unexpected strain *wPip*, which is generally found in the common house mosquito *Culex pipens*.

Armed with low-coverage genomic data from six gomphocerine grasshoppers (including *G. sibiricus*), I carried out a novel comparative analysis of mobile DNA. I found significant proportions of DNA transposons (mean=24%), LINE elements (mean=21%), and satellite DNA (mean=8%). Here, I also observed that satellite DNA and helitron abundance was particularly variable (<1-33%, 7-20%). My main finding suggests that the expansion of satDNA in *G. sibiricus* (genome size=8.9GB) and *Stauroderus scalaris* (genome size=14GB) is a consequence of genome size expansion rather than a cause. I postulate that genome expansion is a multistep process where various phases are governed by varied processes. The novel methods developed for this analysis also allows for new method to estimate

repetitive factions, partition repetitive and non-repetitive sequences of any given genome even with low coverage sequencing data.

Overall, my work provides key resources for genomic and transcriptomic work on a group of organisms that have been challenging because of their unusually large genomes. However, large genomes are not only challenging to work with, but they offer unique opportunities for understanding genome size expansions. My comparative analysis sheds new light on how genome size has evolved in the insect clade with largest genomes.

Zusammenfassung

Genetische Polymorphismen beschreiben die Variation in DNS-Sequenzen zwischen Individuen einer Art oder zwischen Populationen. Die genetischen Polymorphismen zwischen Individuen bis zu ganzen Population und von Punktmutationen zu ganzen Genomen umfassen Milliarden von Basenpaaren. Ein fundamentales Ziel funktioneller Genomik ist es, den Zusammenhang von genetischen Polymorphismen und phänotypischer Variation zu verstehen. Neuste Entwicklungen in der Hochdurchsatzsequenzierung haben es möglich gemacht, diesen Zusammenhang umfassend zu explorieren. Meine Dissertation ergründet genetische und genomische Polymorphismen in Heuschrecken der Unterfamilie Gomphocerinae, einer Insektengruppe mit ungewöhnlich großen und komplexen Genomen, mittels moderner transkriptomischer und genomischer Methoden.

Zu diesem Zweck habe ich eine neue Methode zur Entdeckung von Mikrosatelliten für die Sibirische Keulenschrecke (*Gomphocerus sibiricus*) entwickelt und ein qualitativ hochwertiges Referenztranskriptom erstellt. Dies erleichtert zukünftige Arbeiten mit der Art. Außerdem habe ich gezeigt, wie Hochdurchsatzsequenzierung in Kombination mit bioinformatischen Methoden durch Reduktion der Primerbindungsinterferenzen kosteneffizient zur Entwicklung von Mikrosatelliten eingesetzt werden kann. Durch den wirksamen Einsatz der Informationen, die ich bei dem Transkriptomzusammenbau erhalten habe, konnte ich das gesamte mitochondriale Genom der Art rekonstruieren und die Divergenz der Sequenzen vom mitrochondrialen Genom verwandter Arten und Population abschätzen. Das Assemblierung des Transkriptoms erbrachte auch Hinweise auf das Vorhandensein von *Wolbachia*-Endoparasiten des Stammes wPip, der sonst von der Gewöhnliche Stechmücke *Culex pipens* bekannt ist.

Basierend auf Sequenzdaten mit geringer Sequenziertiefe von sechs Heuschreckarten der Unterfamilie Gomphoerinae (einschließlich *G. sibricus*) habe ich eine neue vergleichende Analyse mobiler DNS durchgeführt. Dabei habe ich signifikante Anteile von DNS-Transposons (25% im Mittel), LINE-Elementen (21%) und Satelliten-DNS (8%) gefunden. Außerdem habe ich festgestellt, dass Satelliten-DNS und Helitrons besonders variabel in ihrer Häufigkeit waren (<1-33% bzw. 7-20%). Meine Ergebnisse deuten darauf hin, dass eine Zunahme von Satelliten-DNS in *G. sibiricus* (Genomgröße = 8.9 GB) und *Stauroderus scalaris*

(Genomgröße = 14 GB) eher eine Folge als eine unmittelbare Ursache der Genomgrößenexpansion sein könnte. Ich postuliere, dass die Genomgrößenexpansion in mehreren Schritten verläuft, bei denen in unterschiedlichen Phasen unterschiedliche Prozesse dominieren. Die für diese Analyse neu entwickelten Methoden erlaubten auch eine neue Abschätzung des Anteils an repetitiver DNS selbst bei geringer Sequenziertiefe.

Insgesamt bietet meine Arbeit neue genomische und transkriptomische Ressourcen zur Arbeit mit einer Organismengruppe, die aufgrund ihrer großen Genome bisher eine große Herausforderung darstellte. Dabei sind die großen Genome nicht nur eine Herausforderung, sondern bieten auch besondere Chancen zum Verständnis von Genomgrößenexpansionen. Meine vergleichende Analyse ermöglicht so neue Einblicke in die Genomgrößenevolution in der Insektengruppe mit den größten Genomen.

Introduction

"What we know as the science of genetics is meant to explain two apparently antithetical observations – that organisms resemble their parents and differ from their parents. That is, genetics deals with both the problem of heredity and problem of variation. It is in fact the triumph of genetics that a single theory down to the molecular level, explains in one synthesis both the constancy of inheritance and its variation. It is the Hegelian dream" – R.C. Lewontin (Lewontin, 1974)

In recent years, with the advent of cheap high throughput sequencing technologies, new avenues of genetic and genomic investigations are now possible. For many decades, most orthopteran species remained stubborn to most genetic investigation tools due to their large, complex and perhaps fascinating genome architecture. Recent cutting-edge methods for *de novo* repetitive element discovery (Goubert *et al.*, 2015a; Novák *et al.*, 2013) have allowed us to explore and understand large and complex genomes. Furthermore, leveraging these new methods have also allowed us to develop and optimize classical genetic methods such as microsatellite markers to be viable again (Shah *et al.*, 2016).

Genetic polymorphism

Genetic diversity is described as the variation in DNA sequence between distinct individuals of a given species or population (Ellegren & Galtier, 2016). Genetic diversity differs considerably among species, loci and even chromosomes. Ellegren and Galtier suggest that this may be theoretically thought as reflecting the balance between the appearance and disappearance of genetic variants, where new alleles appear each generation by spontaneous mutation. Population genetic theory predicts that the genetic diversity of a population increases with increasing number of individuals contributing to reproduction in that population (Hartl & Clark, 2007).

Lewontin's paradox

Under Kimura's neutral theory of molecular evolution, genetic diversity levels at neutral sites reflect the balance between mutational input and the loss of genetic variation due to the random sampling of gametes in a finite population, which under simplistic assumptions

is the rate of genetic drift (Kimura, 1968; Kimura, 1983). However, as natural populations fluctuate in size over time and individuals can greatly vary in their reproductive success, population theory indicates that selection will be more effective in large random-mating populations. Whether this should result in a faster or slower rate loss of genetic variation for a selected site is unclear since some modes of selection lead to loss of genetic variation but others can maintain it (Ohta, 2001). In 1974, Richard Lewontin, observed that while population sizes of different species of fruit flies can vary by a large magnitude, the amount of neutral genetic diversity does not, and appears to bear no simple relationship to effective population size (Lewontin, 1974). More recently, Corbett-Detig, Hartl and Sackton (2015) describe these population genetic theory expectations as succinctly as "under simple neutral models of evolution, levels of neutral genetic diversity within species are expected to increase proportionally with the number of breeding individuals" (Corbett-Detig *et al.*, 2015).

In another recent study, by Romiguier and colleagues (Romiguier *et al.*, 2014), who sampled the transcriptomes of 76 non-model animal species (31 families spread across 8 major animal phyla) evidence was found that long lived or low fecund species with brooding ability were genetically less diverse than short-lived and highly fecund ones. Their results revealed that estimates of nucleotide diversity spanned over 2 orders of magnitude across species and tended to be similar within families but distinct between families. In order to explain the observed variation in nucleotide diversity, they focused on life-history traits. Their analysis suggested that genetic polymorphism is well predicted by species biology whereas historical and contingent factors are only minor determinants of genetic diversity. More succinctly, life history is a major predictor of genetic diversity. However, Corbett-Detig, Hartl and Sackton (2015) argue that natural selection can impact levels of neutral diversity via adaptive fixation of beneficial mutations and selection against deleterious mutations which both purge neutral variants that are linked to selected mutations (Burri, 2017), indicating that when natural selection is ubiquitous across the genome, it can reduce observed levels of neutral polymorphisms (Corbett-Detig *et al.*, 2015).

The evolution of genome size

Genome size varies by several orders of magnitude both within and among diverse taxonomic levels of plants and animals and effects many fitness-related traits such as gene expression, cell size, metabolic rates and body size (Petrov, 2001; Petrov, 2002). Several processes may lead to genome enlargement or reduction. Over evolutionary time these processes lead to clade-specific differences in genome size at higher taxonomic levels as well as distinct variations (Alfsnes *et al.*, 2017). Elliot and Gregory observe the general trend, that broadly across eukaryotic genomes (up-to about 500 Mb in size), there is a linear increase in transposable element (TE) diversity and repetitive content and TEs in particular are the major contributors to genome size (Elliott & Gregory, 2015). However, above this, no clear trend is observed.

The large variation in the genome sizes found across eukaryotes is generally explained through three sets of theories (Gregory, 2018). They all assume that most of this variation is attributed to repetitive and non-genic DNA. First, Lynch and Conery (Lynch & Conery, 2003) proposed that species with large effective population sizes (N_e) will be less likely to tolerate large changes in genome size as the efficacy of selection scales with increasing population size. Therefore, at low population sizes, maladaptive changes may accumulate and persist in the population. This, implies that large changes in genome size can be expected in young species. The second theory of mutational equilibrium, as proposed by Petrov (2001, 2002), suggests that genome size change is gradual and is due to the imbalance between insertions and deletions which eventually reach equilibrium. Hence, some genomes tend to move toward smaller sizes due to higher deletion rates compared to insertion rates and vice versa. Furthermore, this also implies that the effect of insertions and deletions is proportional to the genome size itself. Lastly, the adaptive genome size theory postulates that variation in the amount of non-coding DNA results in significant phenotypic changes and thus evolves under natural selection (Powell 1997). This suggests that genome size variation should track environmental variation as species evolve to exploit habits uniquely (Alfsnes *et al.*, 2017; Kapusta *et al.*, 2017; Yuan *et al.*, 2018).

Lefébure *et al.* (2017) find a substantial correlation between selection efficacy as measured by transcriptome derived d_N/d_S and repeatome size where d_N/d_S is a measure of selective

pressure on amino acid replacement mutation ratio of nonsynonymous to synonymous substitution in the genomes of asellid isopods (Lefébure *et al.*, 2017). This finding indicates that for a large part, genome size is controlled by the efficacy of selection to prevent the invasion of the genome by repeat elements. They propose that although the transcriptome-wide d_N/d_S provides an average estimate of selection efficacy since the divergence of two species of a pair, polymorphism-based proxies are influenced by recent N_e fluctuations, independently of divergence time. However, the authors also suggest that disentangling the forces that drive genome size variation has commonly been complicated by rampant co-variation between genome size and multiple traits such as cell and body sizes, growth rates and metabolism.

Gene duplications may prove to be beneficial by increasing the expression of a gene which may improve fitness, whereas the benefits of mobile DNA or repetitive element accumulation are less clear. However, increasing genome size has a higher physical cost for the individual as the requirement for Nitrogen (N) and Phosphorus (P) is higher as these are major constituents of DNA, and energetically expensive to sequester. A recent study by Guignard *et al.* (2016) shows that N availability is a possible driver of genome size variation in 99 species of *Primulina* (Guignard *et al.*, 2016). Similarly, another study found that the mean genome size of plants grown with N and P fertilizer supplements was significantly higher than those without. Furthermore, with increasing genome size, cell division and metabolism become slower and thereby impede growth and development rates (Cavalier-Smith, 1978; Gregory, 2001). This may affect fitness traits positively or negatively depending on the environment.

Decoding the dynamic mechanisms which regulate genome size is crucial to our understanding of molecular evolution in animals. However, it is quite likely that the processes which govern this phenomenon are diverse. For example, in a prominent study by Kapusta, Suh and Feschotte (2017), which investigated the net change in the amount of DNA across 10 species of eutherian mammals and 24 species of birds, where changes in ploidy are rare and there is no evidence of whole genome duplication events occurring during their recent evolution, with mostly nuclear DNA gains through TE expansion, and little impact on genome size. The authors postulate that the amount of lineage-specific DNA gained by transposition has been adjusted by DNA loss primarily through large-size deletions. They

also suggest that large parts of these genomes are inessential, and this has played an important role in the phenotypic evolution of birds and mammals. They provide specific evidence for metabolic rate and genome size evolution where the metabolic constraints of powered flight keep genomes streamlined.

Interestingly, this trend of genome size and repeatome co-evolution is not observed in squamate reptiles (Pasquesi *et al.*, 2018). Though reptilian genomes are similar to those of birds, they contain highly variable amounts of repeat content, and genome size is highly conserved in this group (Pasquesi *et al.*, 2018). They also report finding no evidence linking genomic repeat abundance or TE activity with that of effective population size. They suggest that the most likely explanation for this is that N_e effects the fixation of deletions leading to a relatively constant genome size in this group.

Recently, Arnquist *et al.* (2015) found that genome size varies marked both between and within seed beetle species and that genome size shows rapid and bidirectional evolution. The pattern of evolution of genome size is not consistent with a major role for genetic drift in shaping genome size, and they found no evidence for correlated evolution with estimated species-specific population sizes. However, they report genome size correlated evolution with reproductive fitness, supporting the hypothesis that genome size variation results from natural selection in this clade. Schielzeth *et al.* (2014) suggest that sexual selection may act to reduce genome size by providing evidence for negative association between song attractiveness and genome size in the grasshopper *Chortippus biguttulus* (Schielzeth *et al.*, 2014).

As flying insects tend to have higher metabolic rates and smaller genome sizes (Reinhold, 1999), a recent study by Lower *et al* (2017) observed a positive relationship between genome size and repetitive DNA, and especially retrotransposons in North American fireflies (*Lampyridae*) (Lower *et al.*, 2017). Here, they also note neither genome size nor significant repeat classes showed evidence of non-neutral evolution and found no evidence of strong selection acting on genome size. They also suggest that genome size evolution is gradual and exhibits complete phylogenetic dependence with no relationship with measured morphological variables.

A cutting edge study Petersen *et al.* (2019) also supports the hypothesis that genome size is correlated with TE content and plays a role in the dynamics of genome size evolution in insects, even though genome size and TE content are highly variable (Petersen *et al.*, 2019). They also report major differences in both TE abundance and diversity between species of the same lineage, and also suggest the possibility of lineage specific similarity in TE elimination mechanisms such as the piRNA pathway, which silences TEs during transcription. Overall, they suggest inter and intra lineage variation in both TE content and composition with a highly variable age distribution of individual TE super-families which indicate a lineage specific burst-like mode of TE proliferation in insect genomes unlike Lower *et al.* (2017).

The role of repetitive DNA and its relation to genome size is complex and perplexing. However, we may observe and measure the major general trends in large genomes with large fractions of repetitive DNA. Here, we could finally tease out if at all there is a relationship between repetitive DNA and genome size. Orthopterans present us with such an opportunity (Figure 1). Most genomic and genetic resources for investigating TEs in orthopterans are scarce outside of agricultural pests. Nonetheless, a handful of emerging methods with new insights into the wonderful world of mobile DNA.

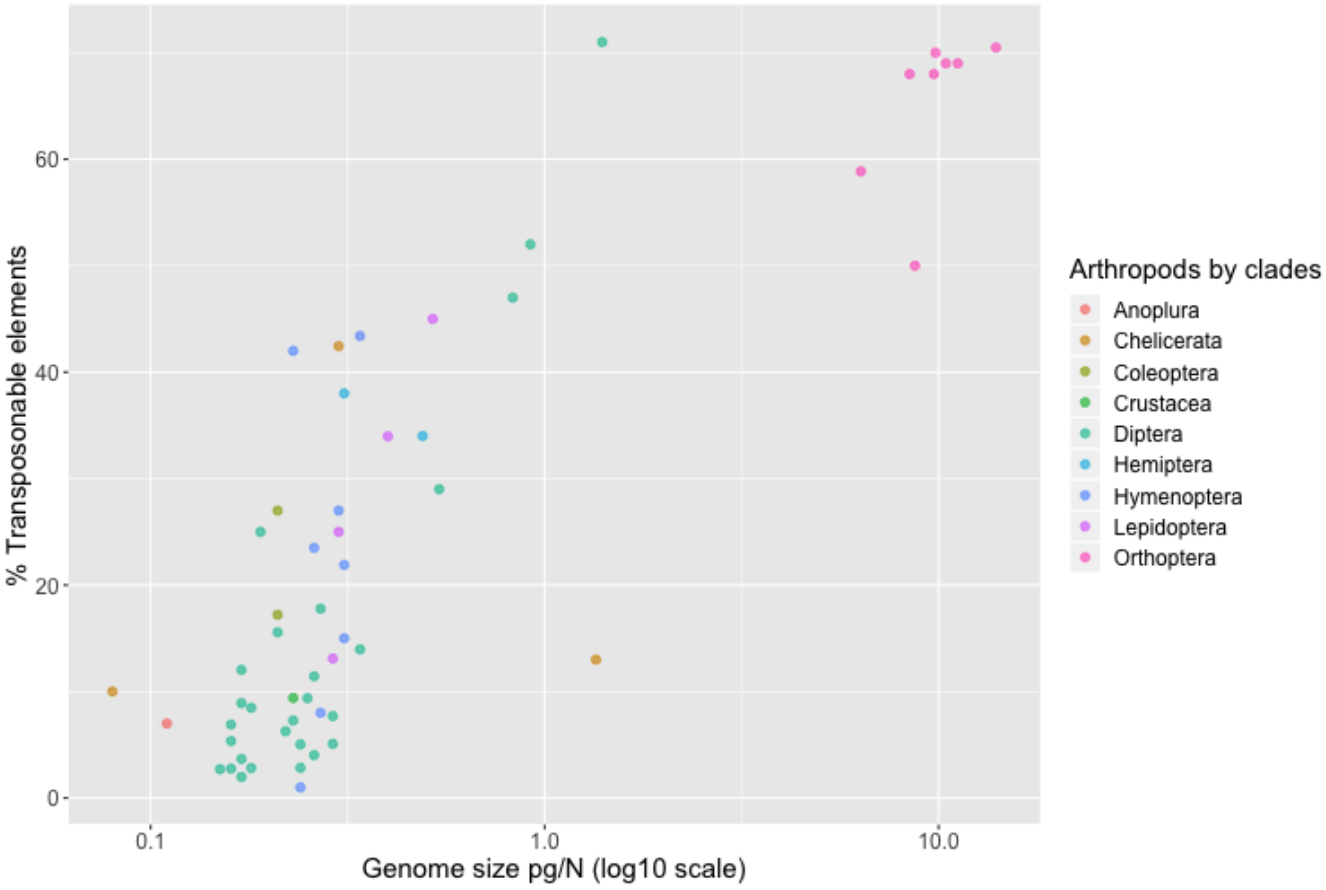


Figure 1: Estimated transposon content to genome size of a few select arthropod clade. Most orthopterans are found in the top right corner, indicating large genome size and a significant proportion of the genome being occupied by transposable elements. Partial data adapted from Canapa et al. (2016).

Computational tools for *de novo* TE and repetitive DNA discovery

Recently, a number of computational methods have been released which can perform *de novo* discovery of TEs which include RepeatExplorer, dnaPipeTE, Tedna, RepARK, REPdenovo and RepLong (Chu *et al.*, 2016; Goubert *et al.*, 2015a; Guo *et al.*, 2017; Koch *et al.*, 2014; Novák *et al.*, 2010; Zytnicki *et al.*, 2014). These methods rely on the high abundance of TE in genomic reads and use low coverage assembly methods to assemble TEs and repetitive DNA, which is usually followed by annotation (Goerner-Potvin & Bourque, 2018). These tools also have the potential to discover new repeat families. However, the task remains daunting as the repetitive fraction of the genome of non-model species remains unknown and these tools are known to be prone to false positives (Caballero *et al.*, 2014). Of these tools, only RepeatExplorer and dnaPipeTE offer *de novo* assembly and annotation of

repetitive DNA assemblies with built-in TE annotation and classification using RepeatMasker (Smit, 2015).

De novo discovery, classification and annotation of transposable elements

RepeatExplorer and dnaPipeTE offer excellent methods to investigate TE in species without extensive repetitive DNA repositories, fully assembled genomes and with low coverage sequencing data.

RepeatExplorer is a full pipeline which efficiently uses a graph-based approach for similarity based partitioning of short read data to construct clusters made of repeats derived from individual repeat families (Novák *et al.*, 2010). This process starts with an all-to-all comparison of read data to find similarities. A modified version of megablast package, which is specifically designed to efficiently perform an all-vs-all search is utilized for this purpose. The results of this search are then used to build clusters of overlapping reads representing different repetitive elements. Due to the low genome coverage of the data, the single-copy sequences are only sparsely covered and seldom overlap, represented by isolated nodes in the graph. However, repetitive sequences represent groups of mutually connected nodes due to frequent sequence overlap. Next, the clusters differentiate between interspersion and partial sequence similarities through graph analysis via a hierarchical agglomeration algorithm which can quantify and characterize individual repeat families. This step prevents merging of multiple distinct elements into the same cluster. Fortunately, this approach is robust to considerable sequence variation in genomic copies of repeated elements. This is followed by cluster analysis which includes graph layout calculation, RepeatMasker (Smit, 2015) search, protein domain search, assembly of cluster contigs and similarity between clusters. Finally, summary HTML reports and cluster contigs are built for the user.

Similarly, dnaPipeTE is a fully automated pipeline designed to assemble and quantify repeats from sequence read data (Goubert *et al.*, 2015a). dnaPipeTE first performs at least three uniform samplings of read data to produce low coverage (<1X) data sets which are subsequently used in the analysis to avoid the assembly of non-repetitive genome content. The first two samples are used in the assembly step, while the last one is used for quantification step. In the assembly step, contigs are built using the Trinity *de novo*

transcriptome assembler (Haas *et al.*, 2013) as the authors assume that similar to transcripts, TE copies from the same family can be observed to accumulate mutations, insertions, deletions and other structural changes, which can be recovered in a similar fashion. Hence, the authors aim to recover a consensus sequences of TE families through this procedure. In subsequent runs, dnaPipeTE adds reads mapping to '*k-mer* contigs' associated with repeats to the second independent sample, which allows for the recovery of more and larger contigs. Next the contigs are annotated with RepeatMasker with a built-in repeat library. Next, BLASTn (Altschul *et al.*, 1990) is used match the third subsample to the annotated contigs assembled by dnaPipeTE and the repeat library. In a further BLASTn step, unmapped reads are matched against unannotated dnaPipeTE contigs. Finally divergence is computed between the dnaPipeTE contigs and the first BLASTn search as a proxy for the TE family.

The *Acrididae* Grasshoppers

Orthopterans provide for an ideal model system for exploring genome size evolution. This insect order contains over 28,000 species described (Cigliano *et al.*, 2018). They have the largest and most variable of all insect genomes with C-values ranging from 1.55 in the cave cricket, *Hadenoecus subterraneus*, to 16.93 in the large mountain grasshopper, *Podisma pedestris* by a significant degree (see Figure 2).

Orthopterans are also crucial to many food webs (Laws *et al.*, 2018). Several species of grasshoppers are considered pests when they develop into local and large-scale population outbursts. Although most grasshoppers seem to have limited dispersal capability due to short wings, they are rather diverse in terms of size, body morphology, ecology and life history traits. In many studies, they can serve as bioindicators as they are ecologically sensitive and mobile, especially in the effect of chemical pollution (Andersen *et al.*, 2001; Devkota & Schmidt, 2000).

Acrididae grasshoppers are the prominent ubiquitous herbivores of grasslands around the world, where they contribute to more than half the above ground arthropod biomass (Branson *et al.*, 2006; Laws *et al.*, 2018). The Acrididae family of grasshoppers includes more than 6700 species and represents one of the most diverse in the sub-order of Caelifera. The Acrididae family had a monophyletic origin in South America in the mid to late Cenozoic era

(Song *et al.*, 2018). Even though the Acrididae is a relatively young group, a recent study by Song *et al* (2018) suggests that their cosmopolitan distribution was achieved by rapid radiation though multiple migrations out of South America to Africa and Eurasia (Song *et al.*, 2018). Within this family, the clade consisting of Acridinae, Gomphocerinae, and Oedipodinae represents the largest subgroups within the Acrididae with over 2700 species (Cigliano *et al.*, 2018). Gomphocerinae, with 192 genera, and 1273 species, is the largest sub-family within the Acrididae. There is evidence that it is a paraphyletic group intermingled with Acridinae and Oedipodinae (Cigliano *et al.*, 2018). However, this group is also one of the best studied for its mating behavior. Here, many gomphoceminae species show pre-copulatory courtship behavior using complex acoustic, vibrational and visual signals.

A large number of grasshopper species occur at least two different colorations in their life. Variable coloration is associated with geophilous, grassland, temperate and alpine habits (Uvarov, 1977)(see Figure 3). The variable coloration of Acrididae grasshoppers is either environmental (for example: *Schistocerca americana*) (Tanaka, 2004) or genetically determined (for example: *Gomphocerus sibiricus*, *Pseudochorthippus parallelus*) (Köhler *et al.*, 2017; Valverde & Schielzeth, 2015). One remarkable color polymorphism is the one of green and brown, which is frequent among many orthopteran species in central and western Europe(Bellmann & Luquet, 2009). Furthermore, this polymorphism is also common in the orders of Phasmatodea and Mantodea which have diverged about 250 MYA from orthoptera (Misof *et al.*, 2014). The green color in orthoptera is formed by tetrapyrroles such as biliverdin, where different oxidation states leads to pigments of yellow, red, violet, blue or green (Rowell, 1971).

Recently, Köhler, Samietz and Schielzeth studied altitudinal variation in morphology and color in the color-polymorphic meadow grasshopper, *Pseudochorthippus parallelus*, which suggested a gradient of morph population proportions, where lowland populations were dominated by the green morph and high-altitude populations by the brown morph (Köhler *et al.*, 2017). This evidence is strongly indicative of local adaptation along an altitudinal gradient with a role in efficient thermoregulation under high-altitude conditions.

In another recent study by Dieker et al 2018, which surveyed 42 sites across the alpine range of *Gomphocerus sibiricus*, observed co-occurrence of green and brown morphs with proportions ranging from 0-70% with considerable spatial heterogeneity where green individuals tended to increase with decreasing summer and winter precipitation (Dieker *et al.*, 2018). The authors argue that small scale migration-selection balance and balancing selection may maintain such polymorphic populations.

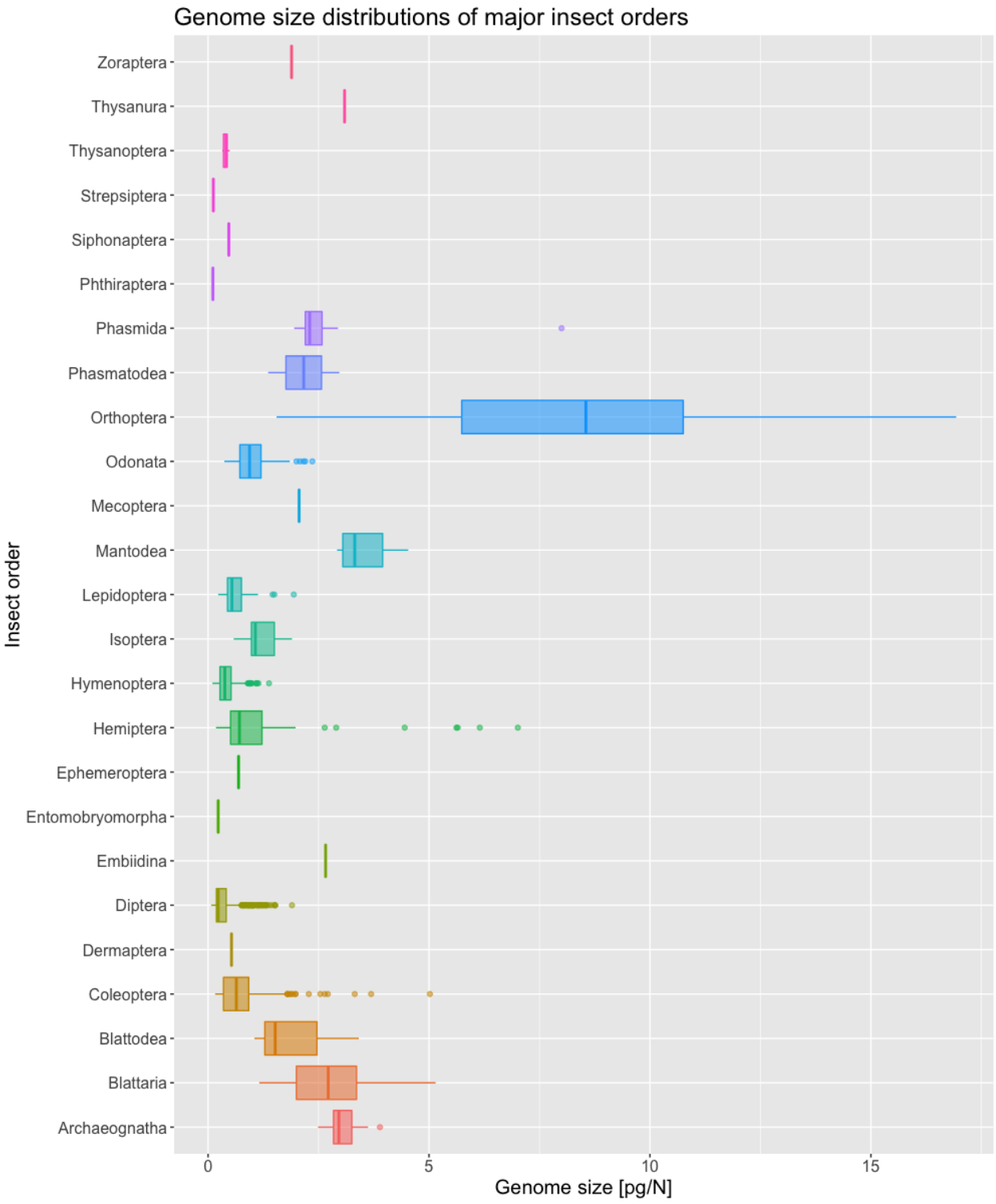


Figure 2: This plot shows the distributions of genome sizes across major insect orders. The order Orthoptera shows the widest variation by a large margin.

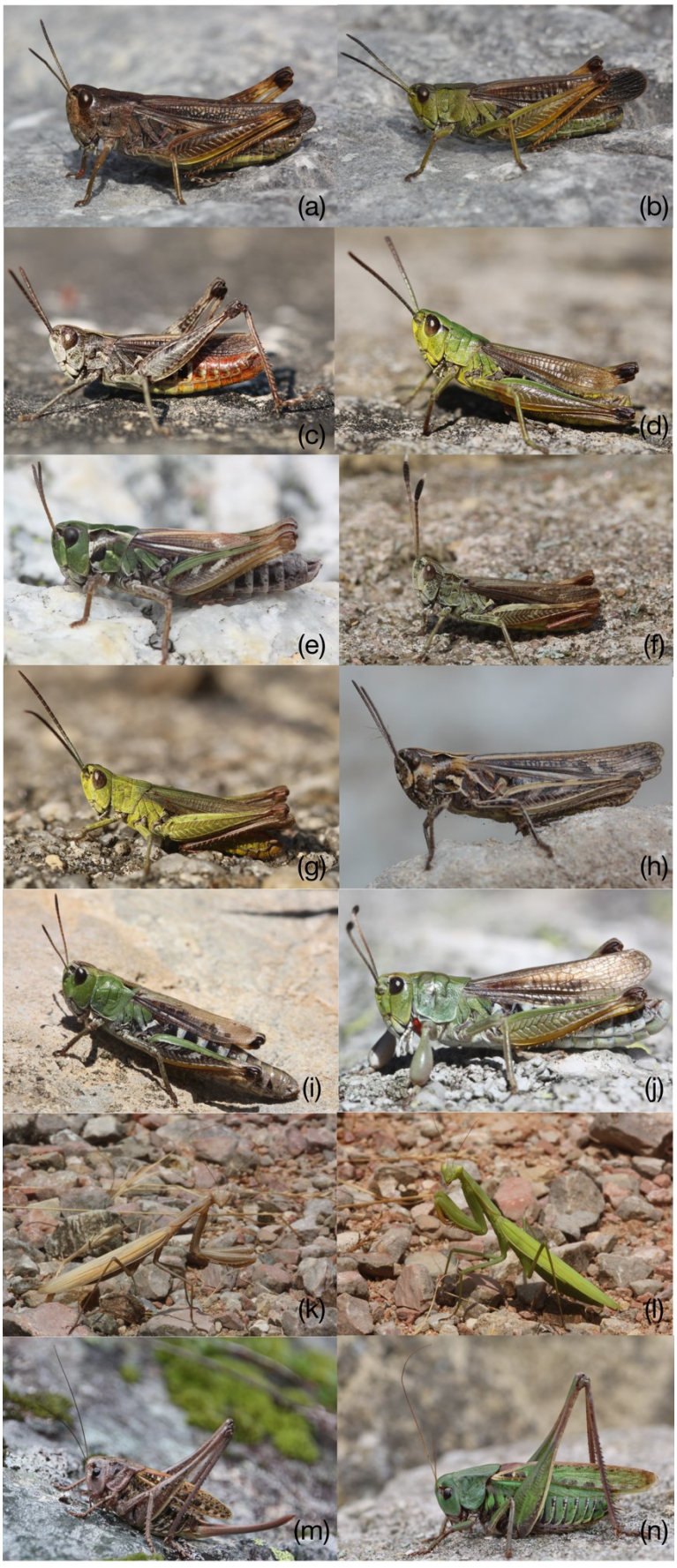

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)

Figure 3: Widespread green brown dimorphism is observed in many members of the polyneoptera clade. (a-b) Brown and green morph of the large mountain grasshopper (Stauroderus scalaris), (c) brown male morph of orange-tipped grasshopper (Omocesthus haemorrhoidalis) (d) meadow grasshopper *(Pseudochorthippus parallalus), (e) green female morph of Aeropedellus variegatus, (f) brown male morph of Gomphocerus rufus, (g) green male morph of* steppe grasshopper *(Chorthippus dorsatus), (h) brown morph of bow-winged grasshopper (Chorthippus biguluttus), (i) female green morph of the club legged grasshopper (Gomphocerus sibiricus),(j) green male G. sibiricus with prominent clubbed legs, (k-l) green and brown morphs of Manta religosa, (m-n) brown and green morphs of the wart-biter bush-cricket (Decticus verrucivorus), a member of the Ensifera group.*

Gene expression analysis and *de novo* transcriptome assembly

A central goal of functional genomics is to establish a link between genetic polymorphism and phenotypic variation. Initially, most gene expression analysis relied on hybridization-based microarray technology which offered limited ability to record various transcripts in a limited dynamic range (Grabherr *et al.*, 2011). However, with the development of high throughput sequencing of cDNA libraries, these limitations were lifted (Wang *et al.*, 2009).

Orthopteran transcriptomes

There have been a few studies utilizing gene expression analysis to investigate orthopterans and till date about 20 orthopteran species have reference transcriptomes assembled. Many of them have been possible through the 1KITE project. However, a few of these have been investigated further mainly for gene expression differences in ontology, phase polymorphism and color polymorphism. In *Chorthippus biguttulus*, key pathways and processes were highlight 9 difference patterns of gene expression which were involved in development pathways that progress embryo to imago (Berdan *et al.*, 2017). This was the first time that the differential gene expression of developmental stages was studied in a hemimetabolous insect. Phase polymorphism from solitarious to gregarious phases, and the associated behavioral, biological, nutritional, metabolical, physiological, color, reproduction and developmental changes in locusts is a remarkable transformation (Anstey *et al.*, 2009; Heifetz *et al.*, 1996; Sword & Simpson, 2000; Wang *et al.*, 2014; Wu *et al.*, 2012). *Locusta migratoria* and *Schistocerca gregaria* are most prominent locust study systems for differential gene expression of phase polymorphism in Orthoptera (Bakkali & Martín-Blázquez, 2018; Jiang *et al.*, 2012). In a recent transcriptomic study of *Schistocerca gregaria* by Bakkali and Martín-Bláquez, suggested that gregarious phase locusts are more active and

exposed to a more stimulating and challenging environment than solitarious ones (Bakkali & Martín-Blázquez, 2018). Furthermore, they also report a positive correlation between population size, behavioral activity and gene expression. In *Locusta migratoria*, Wang and colleagues report that about 28.3% of the gene sets were differentially expressed genes between the brain tissues from locust nymphs experiencing short-term solitarization and gregarization (Wang *et al.*, 2014). Color polymorphism was investigated in *Tetrix japonica* (pygmy grasshopper) using differential expression analysis by Qiu and co-workers (Qiu *et al.*, 2016). They focused on specific putative genes involved in pigment pathways, juvenile hormone, and signaling pathways. More specifically, they identified and probed several pigment related genes involved with melanin, pteridine and ommochome pathways. Out of the 82 pigment genes in their screen, they report that were 12 differentially expressed (Qiu *et al.*, 2016).

Transcript analysis using RNA-seq

Gene expression sequencing by RNA-seq combines transcript discovery and quantification of gene expression into one high throughput sequencing assay (Wilhelm & Landry, 2009). For sequencing, mRNA molecules are converted into cDNA as they provide equivalent sequence information. Current generation of cDNA sequencing technologies allow us to obtain copious amounts of transcriptomic data, in principle, allowing us to identify all expressed, contiguous mRNA sequences for multiple alternatively spliced isoforms and variants, allele-specific expression, and promoters (Bryant *et al.*, 2017). This provides for a cost-effective way to obtain transcriptome data from various species and tissues. RNA-seq can also be used in combination with other biochemical assays to investigate other types of RNA biology (Wang *et al.*, 2009).

In a typical eukaryotic cell, ribosomal RNA (rRNA) constitutes over 90% of the total RNA, while messenger RNA (mRNA) comprises only 1-2%. In order to control for the relatively high abundance of rRNA, two approaches are commonly used. First, rRNA depletion is carried out using selective hybridization of oligonucleotides to rRNA, followed by recognition with a hybrid-specific antibody and subsequent removal of this antibody-hybrid complex on magnetic beads (O'Neil *et al.*, 2013). Second, enrichment of mRNA can be performed using the ability of poly(A)+ tails of mRNA to form stable oligo(dT) under high salt

conditions. Hybrids molecules are destabilized when salt is absent, which allows for mRNA to be recovered (Qing *et al.*, 2013). The choice for mRNA enrichment using poly(A) selection or deplete rRNA depends on the specific experimental design and sample quality as the former requires samples with minimal degradation and yields a higher overall fraction from exonic regions while the latter is preferable when optimal samples are not available. Furthermore, poly(A) mRNA selection enrichment procedure will miss capturing non-coding RNAs. During the cDNA library preparation phase, the extracted RNA is fragmented. This creates a significant hurdle in the accurate *de novo* assembly of novel transcript data. However, this impediment can be overcome by using the outputs of multiple *de novo* assembly software packages followed by removing redundant transcripts via greedy sequence clustering steps (Cerveau & Jackson, 2016). The resultant yields coding transcriptome assemblies that are more credible and perhaps more accurate (Cervea and Jackson 2016). Furthermore, these optimizations also improve the ability to annotate the transcriptome assembly, especially for non-model organisms (Conesa *et al.*, 2016).

De novo transcriptome assembly

Short reads (35-500 bp) obtained from high throughput platforms necessitate the need to reconstruct full-length transcripts through transcriptome assembly. *De novo* assembling of full-length transcripts from short reads is computationally challenging for a number of reasons. Firstly, coverage of transcripts is not uniform, even along the length of same transcript, and often this can vary by several orders of magnitude. Secondly, some transcripts maybe more highly abundant than reads originating from lowly expressed transcripts. Lastly, reads from adjacent loci can overlap and fuse to form chimeric transcripts (Grabherr *et al* 2011).

Most *de novo* transcriptome assemblers are based on a graph-based approach to reconstruct transcripts from a broad range of expression levels and then merge contigs to decrease redundancy. This allows for the discovery of novel isoforms and transcripts as well as novel splicing sites.

Three of the most commonly used *de novo* transcriptome assembly packages are Oases, Trinity and SOAPdenovo-trans (Haas et al., 2013; Schulz et al., 2012; Xie et al., 2014). Most

of these short read assembly software packages rely on de Brujin graphs (DBGs) to determine how and which reads should be joined to form contiguous sequences. The earliest (genomic) short read packages incorporating DGBs (for example: Atlas, ARACHNE, Celera, phrap) relied on an overlap-graph-layout-consensus approach where each read represents a node and each detected overlap as an arc between appropriate nodes. Later assemblers adopted a different approach to utilizing DGBs where the data elements were organized around pieces of sequence of k nucleotides (*k-mers*), and reads were mapped as paths going through the graph going from one node to another in order. Most modern *de novo* transcriptome assembly pipelines run the assembler algorithm at different *k-mer* lengths and then merge these assemblies into one. Lower values of k permits for the assembly of more sensitivity by increasing the connectivity of the graph, leading to the probability of observing overlaps between reads, whereas higher values allow for more specific assemblies, leading to more unique contigs assembled (Cerveau & Jackson, 2016; Conesa et al., 2016).

Overview of the Oases assembler

The Oases transcriptome assembler combines the use of multiple *k-mers* and topological analysis with dynamic error removal adapted for RNA-seq data, and finally merges single *k-mer* assemblies (Schulz *et al.*, 2012). Here, the initial phases of creating a read hash table and the graph structure are identical to these steps in Velvet genome assembler (Zerbino & Birney, 2008). A hash table is an efficient table lookup structure which maps keys to values of the table, and where the hash function computes a binned index, an efficient look-up method. When a *k-mer* is observed in a set of reads, the hash table is used to record the observation with the read ID and position of the *k-mer*. The reads are then converted into a set of original *k-mers* combined with overlaps from previously recorded reads. Next, another database with opposite information is created, where the original *k-mers* are over-lapped by subsequent reads. This ordered set of original *k-mers* is cut each time an overlap with another read begins or ends where the uninterrupted sequences of the original *k*-mers form the nodes, proceeding from one node to the next creates an arc in the graph structure.

Dynamic contig error correction is carried out by a modified TourBus algorithm, which searches through the graph for parallel paths that have the same start and end nodes

(Schulz *et al.*, 2012). If highly similar sequences are found, then the path with lower coverage is merged with the path of higher coverage. This controls for the high variation observed in transcript coverage depth. Furthermore, Oases borrows local the edge removal algorithm from the Trinity *de novo* transcriptome assembly package to remove errors in high coverage regions where the same errors are likely to reoccur. Next, the scaffold is constructed using connection weights which are estimated using the number of spanning reads and the likelihood of observing a read at that position in the contig assuming that to be a normal insert length distribution bound. This is followed by applying multiple static and dynamic coverage filters to the conitgs, as coverage does not reflect the uniqueness of a sequence.

Long contigs, which have a higher likelihood of being unique are then clustered into connected components under the assumption that such components probably belong to the same gene. These are further extended by adding short nodes to the long nodes in the cluster. Redundant long distance connections are removed using transitive reduction to improve efficiency. Branching on the graph most likely indicates alternate splicing events, hence full length isoform transcripts are extracted using information from the topology of the loci. Furthermore, some topologies may not be trivially decomposed, hence a robust heuristic method which uses partial order multiple sequence alignment graphs to resolve them. Finally, as the DBG is highly sensitive to the *k-mer* value selected, the *Oases* pipeline allows for merging of transcript fragments from multiple assemblies (with a range of *k-mer* values). These are added to the graph after the TourBus algorithm removes small or identical transcript fragments.

Overview of Trinity assembler

The Trinity *de novo* transcriptome assembler is perhaps the most commonly used *de novo* transcriptome assembly pipeline available and has been used in thousands of studies and many analysis pipelines (Haas *et al.*, 2013). Trinity uses a novel DBG approach with the assumption that during the assembly process, many disconnected individual graphs would be detected, each representing transcriptional complexity at non-overlapping loci. Hence, Trinity can partition many individual graphs and processes them independently and extracts various isoforms and transcripts. This method is more efficient than computing a full graph

from all reads and an intermediate output of contigs which are strongly supported by many *k-mers*. This process proceeds in three phases, named *Inchworm*, *Chrysalis* and *Butterfly*.

In the Inchworm phase, Trinity uses a greedy *k-mer* based approach for transcript assembly in six steps. First, it constructs a *k-mer* dictionary from all sequenced reads. Second, it removes likely error-containing *k-mers* from the dictionary. Third, it selects the most frequent *k-mer* in the dictionary to seed a contig assembly excluding highly repetitive and singleton *k-mers*. Fourth, it extends the seed in each direction by finding the highest occurring *k-mer* with *k*-1 sequence overlap with current contig terminus and concatenating its terminal base to the growing contig sequence. Trinity then removes the *k-mer* used from the dictionary. Fifth, it extends the sequence in each direction and reports the linear contig. Lastly, it repeats steps three to five till all *k-mers* in the dictionary are depleted.

In the *Chrysalis* stage, Trinity clusters minimally overlapping intermediate contigs obtained at the end of the Inchworm stage into sets of connected components and constructs complete DBGs for each component. Each component defines a collection of contigs that are likely to be derived from alternate splicing isoforms or close paralogs. This is achieved in three steps. First, it recursively groups Inchworm contigs into connected components. Next, it builds a DBG for each component using a word size of *k*-1 to represent nodes and *k* to define the edges connecting the nodes. Lastly, it assigns each read to the component with which it shares the largest number of *k-mers*.

In the *Butterfly* stage, *Trinity* reconstructs plausible full-length linear-transcripts by reconciling the individual DBGs generated by *Chrysalis* with the original reads and paralog genes. It also reconstructs distinct transcripts for splice isoforms and paralogs and resolves ambiguities stemming from errors or from shared sequences between transcripts.

Overview of SOAPdenvo-trans transcriptome assembler

SOAPdenovo-trans (Short Oligonucleotide Analysis Package–transcriptome) (Xie *et al.*, 2014) is another DBG based *de novo* transcriptome assembler, which is derived from SOAPdenovo2 genome assembler (Luo *et al.*, 2012). Unlike SOAPdenovo2, SOAPdenovo-trans integrates the error-removal module from *Trinity* and the robust heuristic graph transversal method from Oases pipelines. SOAPdenovo-trans has two main phases of

assembly, namely contig assembly and transcript assembly. During the contig assembly phase, SOAPdenovo-trans constructs a sparse DBG in exactly the same way as SOAPdenovo2 where the reads are cut into *k-mers* and a large number of linear unique *k-mers* are combined as a group instead of being stored independently. Furthermore, sequencing errors are also detected and removed during this phase. This involves removal of low frequency *k-mers* edges arcs and tips based on a weak depth cut off. Next, the error-removal method from the Trinity package is deployed which filters based on a proportional threshold of maximal depth of adjacent graph elements. In the transcript assembly phase, first reads are mapped back onto the contigs to build linkages. The number of reads is then used to assign weights to these linkages and insert-sizes are used to estimate the distance between them. Next, very short contigs are removed (default contigs below 101 bp), which generally also removes many ambiguous contigs with repetitive content, followed by stringent linearization. In the third stage, the robust heuristic graph transversal method from Oases is used to transverse the graphs generated by clustering contigs into sub-graphs according to their linkage to generate transcripts. Finally, in the gap-filling and correction stage, a DBG-based method which considers all aligned reads during the previous alignment stages in the pipeline. Here, paired-end information is used to cluster semi-unmapped reads into the gap regions followed by local assembly into consensus sequences. This module is based on the gap closing and filling module of SOAPdenovo2. This final gap closing step can be repeated a few times to improve scaffolding and overall median contig length.

Furthermore, downstream in the analysis pipeline, significant computational resources (high performance compute clusters) are required for the annotation and analysis of the assembled transcripts. Merging non-redundant transcripts from multi-*k-mer* assemblies from three highly proven *de novo* transcriptome assemblers provides for an optimal balance between novel transcript discovery and computational effort for *de novo* transcriptome assembly from a non-model species with a complex genome (Conesa *et al.*, 2016; Vijay *et al.*, 2013).

Challenges and overview

Developing genetic and bioinformatic resources for non-model organism, especially with ones with large and highly repetitive genomes is certainly an uphill challenge. Additionally, it

has long been known that even developing basic genetic markers (such as micro-satellites) for large and highly repetitive genome has been predicament (Garner, 2002) for exploratory studies.

Perhaps, such issues are also the likely causes for dearth of resources and information on many orthopteran species. However, emergent bioinformatics methods, third generation sequencing platforms and novel contemporary approaches may provide plausible solutions to some, if not most of these issues. These resources will also allow us to investigate genetic questions in insects with large and complex genomes and transcriptomes with provide insights into genome size evolution and TE evolution. *Arcididae* grasshoppers could serve as a model system to investigate large and complex genomes.

Current RNA-seq technology platforms can facilitate the transcriptome assembly of complex non-model species. One of the most ambitious project pursuing these goals, is the 1KITE (1K Insect Transcriptome Evolution) project, which aims to assemble over 1000 *de novo* transcriptomes from insect species and pursue evolutionary questions. However, the entire Acrididae clade is only represented by two species (*Melanoplinae podismini* and *Stenobothrus lineatus*), of which one (*Stenobothrus lineatus*) transcriptome has already been published previously. Hence, the current efforts that facilitate development of non-model orthopteran species is vital to further our understanding of orthopteran biology and perhaps, large, complex and highly repetitive genomes.

Objectives and Structure

In this thesis, I explored genomic and transcriptomic architecture in *Acrididae* grasshoppers, a group of organisms with unusually large and complex genomes.

In manuscript I, I optimized microsatellite marker discovery for organisms with large and highly repetitive genomes, to facilitate further studies. More specifically, my objective was to dramatically improve the overall odds of discovering polymorphic microsatellite markers in complex genomes with high repetitive content where classical microsatellite markers discovery yields low results due to primer binding interference.

In manuscript II, I constructed a high-quality reference transcriptome for *Gomphocerus sibiricus*, a member of the Acrididae, in order to facilitate the development of this species as a model organism for future studies. Specifically, my objective is to assemble a highly annotated reference transcriptome for *Gomphocerus sibiricus,* which is at least as good, if not better, than other published orthopteran reference transcriptomes.

In manuscript III, my objective was to conduct comparative *de novo* repetitive DNA discovery and analysis in 6 gomphocerine members. My specific objective was to investigate the relationship between genome size and repetitive DNA, with emphasis on satellite DNA and the distribution of major TE families. Furthermore, investigate if there are any major sex-related genomic differences.

Manuscript Overview

Manuscript I

Title: High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome

Authors: Abhijeet Shah, Holger Schielzeth, Andreas Albersmeier, Joern Kalinowski and Joseph Hoffman

Accepted for Publication on: 31th May 2016

Journal: Ecology and Evolution

Supplementary materials: Available online

DOI: 10.1002/ece3.2305

Author contributions: Experimental design, concept and manuscript draft: Abhijeet Shah, Holger Schielzeth, Joe Hoffman; DNA sequencing: Andreas Albersmeier and Joern Kalinowski; bioinformatics analysis: Abhijeet Shah;

Abhijeet Shah: Experimental design, concept and manuscript draft: 40%; bioinformatics analysis: 100%

Manuscript II

Title: Transcriptome assembly for a colour- polymorphic grasshopper (*Gomphocerus sibiricus*) with a very large genome size

Authors: Abhijeet Shah, Joseph Hoffman and Holger Schielzeth

Accepted for Publication on: 30th April 2019

Journal: BMC Genomics

Supplementary materials: Available online

DOI: 10.1186/s12864-019-5756-4

Author contributions: Conceived, designed and manuscript draft: Abhijeet Shah, Holger Schielzeth, Joe Hoffman; Bioinformatic analysis and assembly: Abhijeet Shah

Abhijeet Shah: Conceived, designed and manuscript draft: 50%; bioinformatics analysis and assembly: 100%

Manuscript III

Title: Comparative analysis of genomic repeat content in gomphocerine grasshoppers reveals expansion of satellite DNA and helitrons in species with unusually large genomes

Authors: Abhijeet Shah, Joseph Hoffman and Holger Schielzeth

Author contributions: Conceived and designed: Abhijeet Shah, Holger Schielzeth, Joe Hoffman; Bioinformatic analysis: Abhijeet Shah; manuscript draft: Abhijeet Shah, Holger Schielzeth

Abhijeet Shah: Conceived and designed: 60%; bioinformatics analysis: 100%; manuscript draft: 40%

Manuscript I

Title: High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome

Authors: Abhijeet Shah, Holger Schielzeth, Andreas Albersmeier, Joern Kalinowski and Joseph Hoffman

Ecology and Evolution

Open Access

High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome

Abhijeet B. Shah[1,2], Holger Schielzeth[2,3], Andreas Albersmeier[4], Joern Kalinowski[4] & Joseph I. Hoffman[1]

[1]Department of Animal Behaviour, Bielefeld University, Postfach 100131, 33501 Bielefeld, Germany
[2]Department of Evolutionary Biology, Bielefeld University, Morgenbreede 45, 33615 Bielefeld, Germany
[3]Department of Population Ecology, Institute of Ecology, Friedrich Schiller University Jena, Dornburger Str. 159, 07743 Jena, Germany
[4]Center for Biotechnology, Universitätsstraße 25, 33615 Bielefeld, Germany

Keywords
Acrididae, genetic marker development, *Gomphocerus sibiricus*, high-throughput sequencing, microsatellite, Orthoptera, transposable elements.

Correspondence
Abhijeet B. Shah, Department of Animal Behaviour, Morgenbreede 45, 33615 Bielefeld.
Tel: +49 521 106 2725;
Fax: +49 521 106 2998;
E-mails: ashah@cebitec.uni-bielefeld.de; abhijeet.shah@gmail.com

Funding Information
Deutsche Forschungsgemeinschaft, (Grant/Award Number: "SCHI 1188/1-1") Marie Curie FP7-Reintegration Grant, (Grant/Award Number: "PCIG-GA-2011-303618").

Received: 16 February 2016; Revised: 30 May 2016; Accepted: 31 May 2016

doi: 10.1002/ece3.2305

Abstract

Despite recent advances in high-throughput sequencing, difficulties are often encountered when developing microsatellites for species with large and complex genomes. This probably reflects the close association in many species of microsatellites with cryptic repetitive elements. We therefore developed a novel approach for isolating polymorphic microsatellites from the club-legged grasshopper (*Gomphocerus sibiricus*), an emerging quantitative genetic and behavioral model system. Whole genome shotgun Illumina MiSeq sequencing was used to generate over three million 300 bp paired-end reads, of which 67.75% were grouped into 40,548 clusters within RepeatExplorer. Annotations of the top 468 clusters, which represent 60.5% of the reads, revealed homology to satellite DNA and a variety of transposable elements. Evaluating 96 primer pairs in eight wild-caught individuals, we found that primers mined from singleton reads were six times more likely to amplify a single polymorphic microsatellite locus than primers mined from clusters. Our study provides experimental evidence in support of the notion that microsatellites associated with repetitive elements are less likely to successfully amplify. It also reveals how advances in high-throughput sequencing and graph-based repetitive DNA analysis can be leveraged to isolate polymorphic microsatellites from complex genomes.

Introduction

Although SNPs are increasing in popularity, microsatellites remain an important class of molecular marker due to their low cost and flexibility (Schlotterer 2004). In particular, high levels of polymorphism make microsatellites ideally suited to parentage analysis, particularly for breeding designs involving large numbers of offspring but relatively few candidate parents (Jones and Ardren 2003). In these situations, a handful of highly polymorphic markers can provide a straightforward and cost effective means of constructing pedigree relationships. High levels of polymorphism also make microsatellites suitable for quantifying levels of inbreeding, at least in some species where moderate numbers of microsatellites have been found to outperform substantial panels of SNPs (Forstmeier et al. 2012).

Arguably, one of the greatest disadvantages of microsatellites is the laborious, time consuming and expensive process of developing them in nonmodel species, which until recently required the construction of enriched genomic libraries followed by cloning, hybridization to detect the positive clones and Sanger sequencing (Zane et al. 2002). However, the advent of high-

throughput sequencing approaches, initially Roche 454 but later Illumina sequencing, has simplified the discovery process and now allows many thousands of microsatellite containing sequences to be isolated from virtually any organism (Abdelkrim et al. 2009; Santana et al. 2009; Rico et al. 2013).

Despite the growing ease and popularity of mining for microsatellites *in silico*, a number of issues remain unresolved. In particular, it is still necessary to design oligonucleotide primers from microsatellite flanking sequences and test these for polymorphism in a representative sample of individuals, a process that is both time consuming and costly. Moreover, success rates vary considerably among species (McInerney et al. 2011) and it is not unusual for a significant proportion of primers either to fail to generate interpretable PCR products or to amplify microsatellites that are monomorphic, show evidence of null alleles, or which are inconsistent with a single Mendelian locus (David et al. 2003).

Species with large and complex genomes, including many plants and invertebrates are particularly problematic (Garner 2002). This is because cryptic repetitive elements including transposable elements are disproportionately abundant in large genomes, reaching frequencies as high as 80% in some grasses (Feschotte et al. 2002). Moreover, microsatellites are often not randomly distributed throughout genomes, but instead tend to be preferentially associated with transposable elements such as short interspersed repeats (SINEs) and long interspersed elements (LINEs) (Ramsay et al. 1999). It has even been suggested that repetitive elements could be involved in the genesis and propagation of microsatellites (Arcot et al. 1995; Nadir et al. 1996; Wilder and Hollocher 2001) although it is also possible that transposable element insertion could be favored at sites containing pre-existing microsatellites (Ellegren 2004). Regardless of their exact provenance, microsatellites associated with repetitive elements will exist in multiple copies in the genome where they will have similar or near identical flanking sequences (Zhang 2004). This has been invoked as an explanation for the poor success rates (ranging from zero to around twenty percent) of efforts to develop microsatellites in species as diverse as Norway spruce (Pfeiffer et al. 1997), butterflies (Meglecz et al. 2004), squat lobsters (Bailie et al. 2010), and parasitic nematodes (Grillo et al. 2006).

One way to circumvent this problem is to develop microsatellites from expressed sequence tag libraries (Grillo et al. 2006) or other transcriptomic resources (Blondin et al. 2013), as cryptic elements should be less abundant in selectively constrained regions of the genome. As long as multiple individuals are used for sequencing the transcriptome, this additionally allows microsatellites to be screened for polymorphism *in silico* (Hoffman and Nichols 2011). However, generating a transcriptome is less straightforward than shotgun genome sequencing and also tends to yield microsatellites with lower average levels of polymorphism (Dufresnes et al. 2014). Thus, an attractive alternative would be to identify and remove repetitive sequences from a pool of genomic sequence reads, allowing development efforts to be focused on single-copy microsatellites.

Orthopterans are a group of organisms for which microsatellite development can be particularly problematic. Many species of grasshoppers and locusts are famous for their large genomes (Gregory 2015), like the Acridid grasshoppers, which have haploid genome sizes of around 6–16 Gb (Gregory 2015). A further peculiarity of grasshoppers is the frequent occurrence of facultative (supernumerary) chromosomes that further increase the amount of DNA per cell (Palestis et al. 2004) and hence the potential for primers to bind at multiple sites. The 6.5 Gb genome of the migratory locust *Locusta migratoria* has recently been sequenced and has been found to contain about 60% repetitive elements, of which DNA transposons and LINE retrotransposons are the most abundant (Wang et al. 2014). In addition, a recent study of two other grasshopper species showed by fluorescent *in situ* hybridization that microsatellites are strongly associated with repetitive elements including histone gene spacers, ribosomal DNA intergenic spacers and transposable elements (Ruiz-Ruano et al. 2014).

The number of published microsatellites for Acridid grasshoppers is rather low and these all required the screening of very large numbers of candidate loci (Ustinova et al. 2006; Grace et al. 2009; Chapuis et al. 2012; Keller et al. 2012; Blondin et al. 2013). The club-legged grasshopper, *Gomphocerus sibiricus*, that we study here is an Acridid grasshopper with a sizable genome of around 8.7 Gb (Gregory 2015) and a high prevalence of supernumerary chromosomes (López-Fernández et al. 1986). This species is a valuable model system for studying the evolution sexual ornamentation and the long-term maintenance of color polymorphisms in natural populations (Valverde and Schielzeth. 2015). Fitness assays under competitive conditions in the field and in the laboratory, quantitative genetic studies and inbreeding studies in relation to sexual ornamentation all require genetic markers, yet none are currently available.

Here, we developed an approach for isolating polymorphic microsatellites from complex genomes based on shotgun Illumina MiSeq sequencing and downstream bioinformatic analysis. Specifically, our pipeline incorporates RepeatExplorer (Novak et al. 2013), a collection of software tools that implements graph-based clustering of unassembled sequence reads in order to identify repetitive elements *de novo*. We then exclude reads associated with

clusters of repetitive DNA, identify microsatellite motifs within the remaining singletons using Pal_finder (Castoe et al. 2012) and design primers within Pal_finder using Primer3 (Untergasser et al. 2012). We demonstrate experimentally that primers designed in this way have a significantly greater likelihood of generating clearly interpretable and polymorphic PCR products than primers associated with clusters.

Materials and Methods

Sample collection and preparation for high-throughput sequencing

Grasshoppers were collected near Sierre Valais, Switzerland (46°20′N, 7°30′E) and stored in 70% ethanol at −20°C. Genomic DNA was later extracted from the hind leg using a standard chloroform-isoamyl alcohol protocol (Sambrook et al. 1989).

High-throughput sequencing

Illumina sequencing of five individuals (three males and two females) was conducted at the Center for Biotechnology (CeBiTec) at Bielefeld University. Libraries were prepared with the Nextera DNA Sample Preparation Kit (Illumina, Little Chesterford, UK) according to the manufacturer's instructions. The DNA was then run on a 1.5% agarose gel and fragments in the size range 600–1000 bp were extracted with the Qiagen MinElute Gel Extraction Kit (Qiagen, Hilden, Germany). Fragment sizes were checked using a High Sensitivity DNA Chip on the Agilent 2100 Bioanalyzer (Agilent, Waldbronn, Germany). Quantification was performed using the Quant-iT Picogreen® dsDNA Assay Kit (Life Technologies, Darmstadt, Germany). The libraries were then sequenced on an Illumina MiSeq sequencer using a MiSeq® Reagent Kit v3 (600 cycles; Illumina) to generate 301 bp paired-end reads. FastQ files were generated automatically by the software MiSeq Reporter(version 2.5.1.3: Illumina Inc, 5200 Illumina Way, 92122 San Diego, CA, USA). Analysis within FastQC (Andrews) indicated that the reads were of high quality.

Graph-based repeat characterization and identification

Graph-based clustering and characterization of repetitive sequences was conducted using RepeatExplorer (pipeline version 198+ stable) (Novak et al. 2013) following the developers recommendations (http://repeatexplorer.umbr.cas.cz). Reads that were identified as singletons were retained for microsatellite mining, whereas reads that were assigned to clusters by RepeatExplorer were discarded.

Trinity assembly

As an alternative to using RepeatExplorer to assemble the repetitive DNA elements, we also tested the utility of the *de novo* assembly program Trinity version 2.1.1 (Haas et al. 2013). We first merged the forward and reverse reads within Pear version 0.9.8 (Zhang et al. 2014) using the default parameters. We then assembled the resulting reads within Trinity using the default parameters.

Microsatellite mining and primer design

The resulting singleton reads were mined for microsatellites (more specifically, potentially amplifiable loci or PALs) using the script Pal_finder version 0.02.04 (Castoe et al. 2012). For simplicity and to avoid the issue of primers spanning forward and reverse reads, only the forward reads were used. The reads were interrogated for di-, tri-, tetra-, penta- and hexanucleotides containing at least eight tandem repeats. Within PAL_FINDER version 0.0.2.04, we then used Primer3 version 2.0.0 (Untergasser et al. 2012) to design primers for the target loci. Default parameters were used except for the PCR product size range, which was set to 100–250 bp, and the annealing temperature range, which was set to 55–65°C.

Filtering criteria of PALs

The output from PAL_FINDER was filtered to remove PALs for which primers could not be designed, PALs that occurred in five or more different reads, and PALs where the primer sequences had phred quality scores lower than 29 for at least 95% of the forward and reverse primer bases. As a last step to exclude any loci with multiple copies, we then BLASTed the remaining PALs against all of the forward reads and excluded all PALs with five or more BLAST hits.

In vitro testing of PALs

We attempted to obtain a representative sample of the filtered PALs by randomly selecting 18 of the 484 dinucleotide repeats, 18 of the 78 trinucleotide repeats and all 12 of the tetranucleotide repeats identified above. PCR primer pairs for these loci were tested for polymorphism in eight wild-caught individuals. Each locus was fluorescently labeled using the M13-tail approach (Schuelke 2000) and PCR amplified using a Type It kit (Qiagen). The following PCR profile was used: one cycle of 30 sec at 95°C; 25 cycles of 45 sec at 94°C, 45 sec at 60°C and 45 sec at 72°C; 23 cycles of 30 sec at 94°C, 45 sec at 53°C and 45 sec at 72°C; and one final cycle of 10 min at 72°C. PCR products were resolved by electrophoresis on an ABI 3730xl capillary sequencer.

Pipeline validation

To test whether PALs mined from singletons have greater amplification success than PALs residing within clusters, we additionally evaluated 21 dinucleotide, 21 trinucleotide and six tetranucleotide repeats mined from reads associated with randomly selected clusters. On each PCR plate, we included two positive controls comprising polymorphic loci from the first round of testing.

Multiplexing

We selected 20 polymorphic microsatellites from the first round of testing for incorporation into two PCR multiplexes. These were then used to genotype the original eight individuals of the initial screen plus 32 additional individuals from the same population. For each multiplex reaction, we used a Type It kit (Qiagen) with the following PCR conditions: one cycle of 5 min at 95°C; 25 cycles of 30 sec at 94°C, 90 sec at 60°C, and 60 sec at 72°C; followed by one final cycle of 30 min at 60°C. PCR products were resolved by electrophoresis on an ABI 3730xl capillary sequencer.

Scoring and data analysis

Allele sizes were scored using the program GeneMarker version 2.6_2 (Softgenetics). To ensure high genotype quality, all traces were manually inspected and any incorrect calls were adjusted accordingly. Genepop on the web (Raymond and Rousset 1995) was then used to calculate the observed and expected heterozygosities and to test for deviations from Hardy–Weinberg equilibrium (HWE), specifying a dememorization number of 10,000, 1000 batches and 10,000 iterations per batch.

Ethics statement

All the field samples were taken from individual-rich populations and in accordance with institutional, national, or international legislation and guidelines. No specific permissions were required for the collection of this neither endangered nor protected species outside protected areas.

Results

Three male and two female wild-caught *Gomphocerus sibricus* individuals were sequenced on part of an Illumina MiSeq run, resulting in 3,197,707 paired-end reads totaling approximately 1.92 Gb (approximate average coverage = 0.05 for a genome of 8.7 Gb (Gregory 2015)). These data were subjected to the bioinformatic workflow outlined in Figure 1. First, we used RepeatExplorer to identify and classify repetitive elements (Step *a* of the pipeline in Fig. 1). This program analyzed a subsample of the sequence data comprising 1,372,373 reads. Of these, 929,879 (67.75%) were grouped into 40,548 clusters (Fig. 2), whereas the remaining 442,494 reads were characterized as "singletons". The top 468 clusters, which account for 60.5% of the reads containing repetitive elements were annotated by RepeatExplorer, revealing that a large proportion show similarity to cryptic repetitive DNA elements (Fig. 3). Consequently, in order to improve the probability of success, we discarded reads associated with clusters, leaving only the singleton reads (Step *b*, Fig. 1), from which we mined microsatellites (Step *c*, Fig. 1).

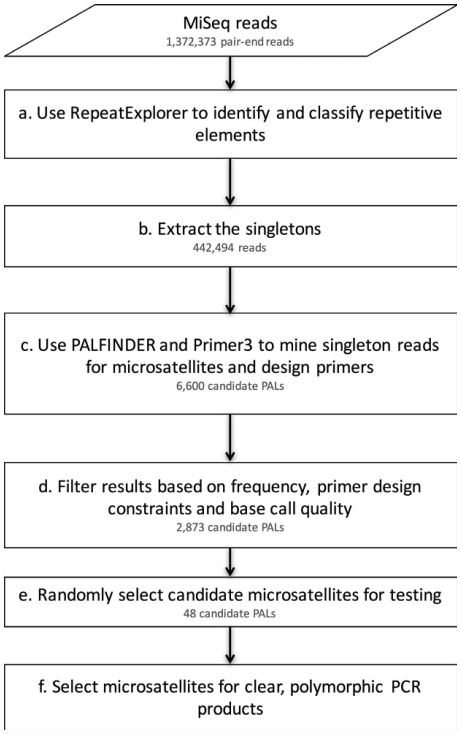


Figure 1. Flowchart detailing the bioinformatic pipeline used to identify polymorphic microsatellites in the club-legged grasshopper, *Gomphocerus sibricus*.

Microsatellite mining and primer design

PAL_FINDER identified 6,600 PALs. Of these, primers could be designed for 2,873 (43.5%) using the parameters

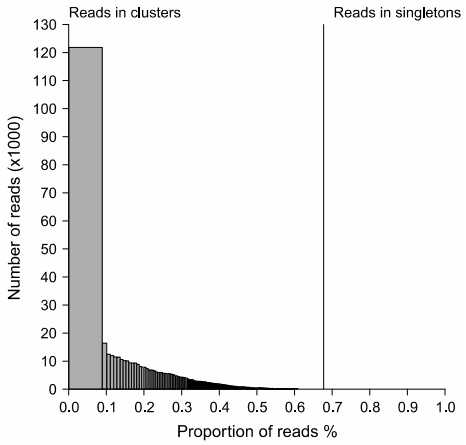


Figure 2. Results of RepeatExplorer analysis showing the numbers of reads classified as clusters or singletons.

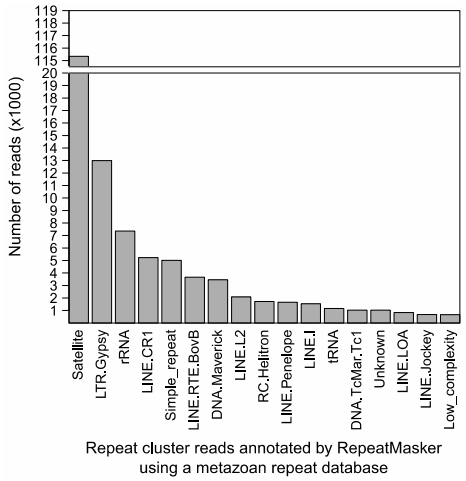


Figure 3. A summary of the RepeatMasker analysis showing the number of reads annotated for repeative elements with at least 550 reads using a metazoan repeat database.

specified in the Materials and methods. PALs with primer sequences that occurred in five or more reads and/or which had phred scores below 29 were removed, leaving 987 PALs (Step *d*, Fig. 1). These carried between eight and 48 tandem repeats (mean = 14.04). We selected 48 of these for *in vitro* testing (Step *e*, Fig. 1).

In vitro verification

Of the 48 primer pairs, 30 (62.5%) yielded clear PCR products that could be discriminated as either polymorphic (n = 29 loci, of which 17 were clearly interpretable and amplified in at least six of the eight individuals) or monomorphic (n = 1 locus) in a sample of eight unrelated *G. sibricus* individuals. The high quality polymorphic loci carried between two and 11 alleles each (mean = five alleles) and observed heterozygosity ranged from 0.125 to 1.0 (Table 1). Four of these loci deviated significantly from HWE (Table 1), although not after false discovery rate (FDR) correction for multiple tests. Of those 18 loci that failed to amplify PCR products resembling microsatellites, six amplified multiple bands that looked similar to amplified fragment length polymorphisms (AFLPs), and the remaining failed to generate any discernable products.

Results of the trinity analysis

Finally, as an alternative to using RepeatExplorer to assemble repetitive DNA elements, we also carried out a *de novo* assembly of the raw reads using Trinity (Haas et al. 2013). This resulted in a total of 153,997 contigs with an N50 value of 677. We then BLASTed the 1223 PALs mined from clusters identified by RepeatExplorer and the 987 PALs mined from singletons against the Trinity assembly using a minimum identity match of 95% and only retaining the top hit. The majority of the PALs mined from RepeatExplorer clusters (908, 74.2%) revealed top hits to the Trinity assembly, while almost none of the PALs residing within singletons showed sequence homology to the Trinity contigs. This indicates that Trinity preferentially assembled the repetitive elements and hence could be used as an alternative to RepeatExplorer.

Pipeline verification

Our microsatellite discovery pipeline is based on the premise that PALs residing within singleton reads should amplify more successfully than PALs residing within clusters. To test this prediction empirically, we evaluated a "control" set of 48 PALs mined from reads assigned to clusters of repetitive elements (see Materials and Methods for details). Only five of these loci (10.4%) yielded polymorphic PCR products consistent with the amplification of a single locus. A further two microsatellites were polymorphic but appear to be duplicated as individuals carry up to four alleles each. Ten additional loci were

Table 1. *In vitro* verification of the primer pairs. Shown are the polymorphism characteristics of 17 microsatellite loci that amplified clearly interpretable and polymorphic PCR products in eight unrelated *Gomphocerus sibricus* individuals.

Locus	Repeat motif	Tandem repeats	Forward primer	Reverse primer	Number of alleles	H_O	H_E	HWE *P*-value
Gsib01	TC	16	AGAGGGAGACAGATAGACGGC	TTCCACACTTTTAAGACTGAATGC	10	1.00	0.93	1.00
Gsib02	TC	10	CTGATTCACAGATAGGGGCG	GTCCATATCCTCCTCCCTCC	5	0.50	0.82	0.09
Gsib07	AC	8	ACACACAACTGCAAACTCCG	TCTTCAGAAAAGATCTCTCCCC	11	1.00	0.93	1.00
Gsib08	TC	8	AGAGACCACAGGCAGAGAGC	CCCTTTATTGATCGCAAAGC	2	0.17	0.53	0.15
Gsib13	TC	21	TGAAATCCATGTAGCATCGC	CGGACTTCAACGAAGATTCC	9	0.88	0.12	0.88
Gsib16	TC	8	TGTGCGATCTTACTCGACCC	GGCCACTTCTTTGTCAGAGC	6	0.38	0.86	0.01
Gsib18	TC	11	AAGGGAGAAGGAAGACGTGC	GAGAAACATGATGTCGACCG	8	0.75	0.91	0.08
Gsib19	ATC	10	TCTATGCTCCAGACGGAACG	CAGACATGAAGCCAAAACCC	6	0.88	0.82	0.57
Gsib21	ATC	9	ACACAAAATATTCCGTGCCC	GACTTACACCAGGTAGGGCG	3	0.50	0.66	0.38
Gsib24	ATC	9	AGTCTAACGGCCAGAAATGC	TAGTTTTGGCGAAGGAGTCG	3	0.75	0.67	1.00
Gsib28	ATT	8	ATGTTCATGGTGACAATGCC	CCCCTCACAGGTTATCTTTGC	2	0.25	0.50	0.22
Gsib29	ATT	8	TCTAGAACCTTTGGTCTGTTGC	ACGAATGTCCCAAGAACAGG	3	0.12	0.61	0.01
Gsib32	TCC	10	CTACCTTCCTCCTATCGCCC	ATGTGGTTCCCTGTTTCTGC	6	0.75	0.84	0.38
Gsib35	ATC	10	TATGCTGCAATATGCTTGGC	TCCTCACAAGTGCAGAATGC	3	0.50	0.42	1.00
Gsib42	ATAC	9	GAGGCTGTAGCCATTTCTCG	GTCTTCACTTCCCATGAGGC	3	0.17	0.71	0.01
Gsib45	AAAG	8	CAAGGCCACAGTTAAGGAGG	AATGTCTGTGAAATATTACGTGCC	3	0.62	0.67	0.66
Gsib46	AATC	9	TATTGCCTCTGAATCTGCCC	ATATAGCTGTCCTCAGCGCC	2	1.00	0.53	0.03

HWE, Hardy–Weinberg equilibrium.

monomorphic and the remaining 31 loci failed to generate interpretable banding patterns. Multiple peaks, often resembling AFLP profiles, were observed in 27 of the latter, while four failed to generate any PCR products. The difference in the success rate of PALs within singletons and clusters, defined as the proportion generating polymorphic banding patterns consistent with a single locus, was highly significant (29/48 versus 5/48; two-tailed binomial proportions test, $\chi^2 = 24.1$, df = 1, $P < 0.0001$).

Multiplexing

Finally, we selected twenty loci for inclusion in two multiplexes and genotyped these in a larger panel of 40 unrelated individuals. One locus did not amplify polymorphic PCR products and a further ten loci deviated significantly from HWE after FDR correction (Table 2). The remaining nine loci were clearly interpretable and did not deviate from HWE.

Discussion

Poor microsatellite amplification success is often associated with the occurrence of cryptic repetitive elements (Pfeiffer et al. 1997; Grillo et al. 2006; Bailie et al. 2010; McInerney et al. 2011). A number of studies reached this conclusion on the basis of *post hoc* analyses of flanking sequence similarities revealed by all-against-all BLAST analysis (Meglecz et al. 2004; McInerney et al. 2011) and through comparison to the Repbase database of known transposable elements (McInerney et al. 2011). Our approach also exploits information on sequence similarity and homology to the Repbase database, but this time through graph-based cluster analysis implemented within RepeatExplorer. However, it differs from previous approaches in two ways. First, we exploited high-throughput sequencing to generate millions of reads, providing greater resolution of the composition of repetitive elements in the grasshopper genome, and second, we conducted the bioinformatic analysis prior to primer design and testing, allowing us to focus on single-copy loci.

Repetitive elements in the club-legged grasshopper genome

Ours is not the first study to assign microsatellite flanking sequences to different families of repetitive element (Bailie et al. 2010; McInerney et al. 2011), although the use of high-throughput sequencing allowed us to scale up from a few hundred Sanger sequences to over three million reads of similar length. By sequencing a library that was not enriched for microsatellite motifs, we could additionally obtain a tentative estimate of the overall proportion of genomic sequences containing repetitive elements. We found that over 67.8% of reads could be grouped into clusters, the top 468 of which accounted for approximately 60.5% of the sequence data. The fraction of repetitive elements is in close agreement with the migratory locust (Wang et al. 2014). However, the club-legged grasshopper is unusual in that a single repeat class cluster

Table 2. Polymorphism characteristics of 20 microsatellite loci that were multiplexed and amplified in 40 unrelated *Gomphocerus sibricus* individuals. The initial PCR mixes (1 and 2) were modified to minimize interference between loci, resulting in mixes 1a, 1b, 2a, and 2b. * denotes HWE tests that remained significant after table-wide false discovery rate correction for multiple statistical testing.

	Mix											
Locus	1	1a	1b	2	2a	2b	Dye	Size range (bp)	Number of alleles	H_O	H_E	HWE *P*-value
Gsib01	x	x					FAM	99–151	22	0.53	0.93	<0.0001*
Gsib02	x	x	x				PET	244–255	10	0.64	0.82	0.0517
Gsib03	x						FAM	280–284	2	0.11	0.10	1.00
Gsib07				x	x		FAM	93–132	19	0.64	0.89	<0.0001*
Gsib09				x	x	x	NED	182	1	1.00	1.00	NA
Gsib13	x	x	x				PET	167–232	17	0.38	0.90	<0.0001*
Gsib14	x	x	x				VIC	119–129	5	0.21	0.71	<0.0001*
Gsib16					x		FAM	188–232	14	0.72	0.86	0.1961
Gsib17	x		x				FAM	180–212	11	0.24	0.68	<0.0001*
Gsib18				x		x	FAM	162–223	22	0.74	0.93	<0.0001*
Gsib19	x						FAM	236–276	12	0.74	0.91	0.0304
Gsib21	x	x	x				VIC	171–186	6	0.36	0.68	<0.0001*
Gsib23		x					FAM	232–253	5	0.29	0.74	0.0001*
Gsib24				x	x	x	VIC	157–169	3	0.44	0.52	0.2274
Gsib28				x	x	x	NED	223–235	5	0.37	0.53	0.0937
Gsib29				x	x	x	PET	234–265	5	0.11	0.61	<0.0001*
Gsib32				x	x	x	VIC	202–230	11	0.67	0.86	0.029
Gsib35	x	x	x				VIC	204–226	3	0.38	0.32	0.706
Gsib36				x	x	x	PET	154–170	8	0.28	0.80	<0.0001*
Gsib45	x	x	x				NED	220–232	5	0.63	0.67	0.2345

HWE, Hardy–Weinberg equilibrium; H_O, observed heterozygosity; H_E, expected heterozygosity.

dominates much more strongly than in other species with large genomes, including animals and plants (Piednoel et al. 2012; Garcia et al. 2015).

In our sample, the most abundant RepeatMasker hit was satellite DNA (see Fig. 3). This is not necessarily surprising as a previous study by Rafferty and Fletcher (Rafferty and Fletcher 1992) found that around 30% of the genome of *Stauroderus scalaris*, another member of the Gomphocerinae grasshopper family, also comprises satellite DNA. With limited data available on other Orthopteran species, we can only speculate as to how and to what extent the distribution and composition of repetitive elements differs among related taxa. The closest relative with genomic resources available, the migratory locust *Locust migratoria*, differs considerably in the composition of repetitive elements, a significant portion of which are DNA transposons and LINE retrotransposons. However, this might not be too surprising because the two species are divergent by approximately 57 million years (Song et al. online early) and even their genome sizes differ considerably (6.5 Gb versus 8.7 Gb, a 33% difference). Exploring how different classes of repetitive element may have invaded the genomes of different Orthopteran species now seems feasible given that more than one species could be pooled onto a single MiSeq run, providing a fertile avenue for future research.

Microsatellite development success

We found that microsatellites developed from singleton reads had a six-fold higher success rate, defined by the proportion of loci amplifying polymorphic products consistent with a single locus, relative to a set of microsatellites mined from reads associated with clusters. This supports a previous study of Norway spruce, which found that primer pairs amplifying a single polymorphic microsatellite were largely restricted to unique clone sequences that lacked repetitive DNA (Pfeiffer et al. 1997). Causes of microsatellite failure in our study included (1) PCR amplification failure, which resulted either in no discernible product or in a small number of nonspecific bands; (2) the amplification of monomorphic bands resembling microsatellite alleles; or (3) the amplification of more than one locus, indicated by the presence of up to four alleles within an individual. All of these patterns have been observed in similar studies of other species with complex genomes (Pfeiffer et al. 1997; Bailie et al. 2010; McInerney et al. 2011).

Elsewhere, microsatellites developed from a nematode species with an apparently complex genome were found to carry unusually high frequencies of nonamplifying or "null" alleles, indicated by the presence of multiple apparently homozygous non-amplifying individuals (Grillo et al. 2006). As null alleles are caused by polymorphisms within the primer binding sites, these authors concluded that the species in question probably has very high levels of sequence polymorphism, reflecting the vast effective population sizes of many nematodes. It is for this reason that we selected twenty loci for inclusion in two mastermixes, which we then used to genotype a larger panel of 40 individuals. Having done this, we found that a considerable proportion of the loci did not conform to HWE in the larger sample. As the majority of these loci showed heterozygote deficiency, we conclude that null alleles may also be relatively common in club-legged grasshoppers. Nevertheless, nine of the loci conformed to HWE, suggesting that with our approach it is eminently feasible to generate a panel of microsatellites large enough for most purposes.

Possible alternatives to RepeatExplorer analysis

In this paper, we focused on using RepeatExplorer to *de novo* assemble and annotate repetitive elements. In principle, however, alternative assembly programs might be used to similar effect. To test this, we also *de novo* assembled our data using Trinity and then looked for overlap between PALs identified in the previous analysis as residing within clusters and singletons, respectively. We found that PALs residing within the clusters identified by RepeatExplorer predominantly mapped to the Trinity assembly. By implication, the Trinity assembly must be enriched for repetitive elements in the same way as the RepeatExplorer clusters, and hence it appears that both approaches could be useful for screening out PALs residing within repetitive elements. It would be interesting in the future to test whether such approaches bring similar benefits in other species and to further explore the merits of other approaches for *de novo* repeat discovery and sequence assembly.

Conclusions

We used massively parallel sequencing together with graph-based clustering and annotation to develop polymorphic microsatellites for the club-legged grasshopper, an emerging quantitative genetic model system. Our study not only sheds light on the composition of the repetitive fraction of this species genome, but also demonstrates the potential of *in silico* filtering to dramatically improve the success of microsatellite development efforts.

Acknowledgments

We are grateful to Elke Hippauf and Amy R. Backhouse for testing the primers for PCR amplification. This research was supported by a Marie Curie FP7-Reintegration Grant within the 7th European Community Framework Programme (PCIG-GA-2011-303618) and an Emmy-Noether fellowship by the German Research Foundation (DFG; SCHI 1188/1-1). We acknowledge support for the Article Processing Charge by the Deutsche Forschungsgemeinschaft and the Open Access Publication Funds of Bielefeld University Library.

Conflict of Interest

None declared.

Data Accessibility

The raw read sequences used in this analysis have been deposited in the short read archive with BioProject ID: PRJNA321244. Details of the microsatellites are provided in the Dryad data repository with the DOI: 10.5061/dryad.23r6v.

References

Abdelkrim, J., B. C. Robertson, J.-A. L. Stanton, and N. J. Gemmell. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46:185–191.

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data *in* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, editor.

Arcot, S. S., Z. Wang, J. L. Weber, P. L. Deininger, and M. A. Batzer. 1995. Alu repeats: a source for the genesis of primate microsatellites. Genomics 29:136–144.

Bailie, D. A., H. Fletcher, and P. A. Prodohl. 2010. High incidence of cryptic repeated elements in microsatellite flanking regions of Galatheid genomes and its practical implications for molecular marker development. J. Crustac. Biol. 30:664–672.

Blondin, L., L. Badisco, C. Pages, A. Foucart, A.-M. Risterucci, C. S. Bazelet, et al. 2013. Characterization and comparison of microsatellite markers derived from genomic and expressed libraries for the desert locust. J. Appl. Entomol. 137:673–683.

Castoe, T. A., A. W. Poole, A. P. J. de Koning, K. L. Jones, D. F. Tomback, S. J. Oyler-McCance, et al. 2012. Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PLoS ONE 7 (2):e30953.

Chapuis, M. P., R. Streiff, and G. A. Sword. 2012. Long microsatellites and unusually high levels of genetic diversity in the Orthoptera. Insect Mol. Biol. 21:181–186.

David, L., S. Blum, M. W. Feldman, U. Lavi, and J. Hillel. 2003. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. Mol. Biol. Evol. 20:1425–1434.

Dufresnes, C., A. Brelsford, P. Beziers, and N. Perrin. 2014. Stronger transferability but lower variability in transcriptomic- than in anonymous microsatellites: evidence from Hylid frogs. Mol. Ecol. Resour. 14:716–725.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5:435–445.

Feschotte, C., N. Jiang, and S. R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. 3:329–341.

Forstmeier, W., H. Schielzeth, J. C. Mueller, H. Ellegren, and B. Kempenaers. 2012. Heterozygosity-fitness correlations in zebra finches: microsatellite markers can be better than their reputation. Mol. Ecol. 21:3237–3249.

Garcia, G., N. Rios, and V. Gutierrez. 2015. Next-generation sequencing detects repetitive elements expansion in giant genomes of annual killifish genus Austrolebias (Cyprinodontiformes, Rivulidae). Genetica 143:353–360.

Garner, T. W. J. 2002. Genome size and microsatellites: the effect of nuclear size on amplification potential. Genome 45:411–419.

Grace, T., A. Joern, J. L. Apple, S. J. Brown, and S. M. Wisely. 2009. Highly polymorphic microsatellites in the North American snakeweed grasshopper, *Hesperotettix viridis*. J. Orthoptera. Res. 18:19–21.

Gregory, T. R. 2015. Animal Genome Size Database http://www.genomesize.com.

Grillo, V., F. Jackson, and J. S. Gilleard. 2006. Characterisation of *Teladorsagia circumcincta* microsatellites and their development as population genetic markers. Mol. Biochem. Parasitol. 148:181–189.

Haas, B. J., A. Papanicolaou, M. Yassour, M. G. Grabherr, P. D. Blood, and J. Bowden. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Hoffman, J. I., and H. J. Nichols. 2011. A novel approach for mining polymorphic microsatellite markers *in silico*. PLoS ONE 6:e23283.

Jones, A. G., and W. R. Ardren. 2003. Methods of parentage analysis in natural populations. Mol. Ecol. 12:2511–2523.

Keller, D., E. Jung, and R. Holderegger. 2012. Development of microsatellite markers for the wetland grasshopper *Stethophyma grossum*. Conserv. Genet. Resour. 4:507–509.

López-Fernández, C., C. G. Delavega, and J. Gosálvez. 1986. Unstable B-chromosomes in *Gomphocerus sibiricus* (Orthoptera). Caryologia 39:185–192.

McInerney, C. E., A. L. Allcock, M. P. Johnson, D. A. Bailie, and P. A. Prodohl. 2011. Comparative genomic analysis reveals species dependent complexities that explain difficulties with microsatellite marker development in molluscs. Heredity 106:78–87.

Meglecz, E., F. Petenian, E. Danchin, A. Coeur D'Acier, J.-Y. Rasplus, and E. Faure. 2004. High similarity between flanking regions of different microsatellites detected within each of two species of Lepidoptera: *Parnassius apollo* and *Euphydryas aurinia*. Mol. Ecol. 13:1693–1700.

Nadir, E., H. Margalit, T. Gaillily, and S. A. Ben-Sasson. 1996. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. Proc. Natl Acad. Sci. USA 93:6470–6475.

Novak, P., P. Neumann, J. Pech, J. Steinhaist, and J. Macas. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 29:792–793.

Palestis, B. G., R. Trivers, A. Burt, and R. N. Jones. 2004. The distribution of B chromosomes across species. Cytogenet. Genome Res. 106:151–158.

Pfeiffer, A., A. M. Olivieri, and M. Morgante. 1997. Identification and characterisation of microsatellites in Norway spruce (*Picea abies* K.). Genome 40:411–419.

Piednoel, M., A. J. Aberer, G. M. Schneeweiss, J. Macas, P. Novak, H. Gundlach, et al. 2012. Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. Mol. Biol. Evol. 29:3601–3611.

Rafferty, J. A., and H. L. Fletcher. 1992. Sequence analysis of a family of highly repeated DNA units in *Stauroderus scalaris* (Orthoptera). Int. J. Genome Res. 1:1–16.

Ramsay, L., M. Macaulay, L. Cardle, M. Morgante, S. D. Ivanissevich, E. Maestri, et al. 1999. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J. 17:415–425.

Raymond, M., and F. Rousset. 1995. Genepop (Version 1.2) - population genetics software for exact tests of ecumenicism. J. Hered. 86:248–249.

Rico, C., E. Normandeau, A.-M. Dion-Cote, M. I. Rico, R. Cote, and L. Bernatchez. 2013. Combining next-generation sequencing and online databases for microsatellite development in non-model organisms. Sci. Rep. 3:3376.

Ruiz-Ruano, F. J., A. Cuadrado, E. E. Montiel, J. P. M. Camacho, and M. D. López-León. 2014. Next generation sequencing and FISH reveal uneven and nonrandom microsatellite distribution in two grasshopper genomes. Chromosoma 124:221–234.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbour Laboratory Press, New York, NY.

Santana, Q. C., M. P. A. Coetzee, E. T. Steenkamp, O. X. Mlonyeni, G. N. A. Hammond, M. J. Wingfield, et al. 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46:217–223.

Schlotterer, C. 2004. The evolution of molecular markers-just a matter of fashion? Nat. Rev. Genet. 5:63–69.

Schuelke, M. 2000. An economic method for the fluorescent labelling of PCR fragments. Nat. Biotechnol. 18:233–234.

Song, H., C. Amédégnato, M. M. Cigliano, L. Desutter-Grandcolas, S. W. Heads, Y. Huang, et al. Online early. 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. Cladistics 31:621–651. doi 10.1111/cla.12116

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, et al. 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res. 40:e115.

Ustinova, J., R. Achmann, S. Cremer, and F. Mayer. 2006. Long repeats in a huge genome: microsatellite loci in the grasshopper *Chorthippus biguttulus*. J. Mol. Evol. 62:158–167.

Valverde, K., and H. Schielzeth. 2015. What triggers colour change? Effects of background colour and temperature on the development of an alpine grasshopper BMC Evol. Biol. 15:168.

Wang, X. H., X. D. Fang, P. C. P. C. Yang, X. T. Jiang, F. Jiang, D. J. Zhao, et al. 2014. The locust genome provides insight into swarm formation and long-distance flight. Nat. Commun. 5:1–9.

Wilder, J., and H. Hollocher. 2001. Mobile elements and the genesis of microsatellites in dipterans. Mol. Biol. Evol. 18:384–392.

Zane, L., L. Bargelloni, and T. Patarnello. 2002. Strategies for microsatellite isolation: a review. Mol. Ecol. 11:1–16.

Zhang, D.-X. 2004. Lepidopteran microsatellite DNA: redundant but promising. Trends Ecol. Evol. 19: 507–509.

Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30:614–620.

Manuscript II

Title: Transcriptome assembly for a colour-polymorphic grasshopper (Gomphocerus sibiricus) with a very large genome size

Authors: Abhijeet Shah, Joseph Hoffman and Holger Schielzeth

Shah *et al. BMC Genomics* (2019) 20:370
https://doi.org/10.1186/s12864-019-5756-4

BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptome assembly for a colour-polymorphic grasshopper (*Gomphocerus sibiricus*) with a very large genome size

Abhijeet Shah[1,2*], Joseph I. Hoffman[2] and Holger Schielzeth[1]

Abstract

Background: The club-legged grasshopper *Gomphocerus sibiricus* is a Gomphocerinae grasshopper with a promising future as model species for studying the maintenance of colour-polymorphism, the genetics of sexual ornamentation and genome size evolution. However, limited molecular resources are available for this species. Here, we present a de novo transcriptome assembly as reference resource for gene expression studies. We used high-throughput Illumina sequencing to generate 5,070,036 paired-end reads after quality filtering. We then combined the best-assembled contigs from three different de novo transcriptome assemblers (Trinity, SOAPdenovo-trans and Oases/Velvet) into a single assembly.

Results: This resulted in 82,251 contigs with a N50 of 1357 and a TransRate assembly score of 0.325, which compares favourably with other orthopteran transcriptome assemblies. Around 87% of the transcripts could be annotated using InterProScan 5, BLASTx and the *dammit!* annotation pipeline. We identified a number of genes involved in pigmentation and green pigment metabolism pathways. Furthermore, we identified 76,221 putative single nucleotide polymorphisms residing in 8400 contigs. We also assembled the mitochondrial genome and investigated levels of sequence divergence with other species from the genus *Gomphocerus*. Finally, we detected and assembled *Wolbachia* sequences, which revealed close sequence similarity to the strain pel wPip.

Conclusions: Our study has generated a significant resource for uncovering genotype-phenotype associations in a species with an extraordinarily large genome, while also providing mitochondrial and *Wolbachia* sequences that will be useful for comparative studies.

Keywords: Insects, Orthoptera, Acrididae, Gomphocerinae, Transcriptome, Mitochondria, *Wolbachia*

Background

One important goal of functional genomics is to establish links between genetic polymorphisms and phenotypic variation [1]. Recent developments in high-throughput sequencing technologies have greatly facilitate this endeavour. However, there are still major challenges inherent to genomic approaches for genotype-phenotype associations in non-model organisms [2]. One of these challenges is imposed by species with large genomes, as genome assemblies are difficult to construct for highly repetitive regions [3]. To some degree, this issue will be mitigated by the development of long-range sequencing technologies [4]. Nevertheless, we expect that genomic approaches for taxa with very large genomes will remain challenging for some time.

Transcriptomics offers an alternative to genomic approaches, as the size of the transcriptome does not scale linearly with genome size and most functional differences with phenotypic effects should be reflected in the transcriptome [5]. Not only do transcripts differ in their sequences, mirroring underlying coding DNA sequence differences, but quantitative analysis can also allow the assessment of regulatory variation influencing transcript abundance [6, 7]. One prerequisite for an analysis of transcript abundance is a high quality transcriptome assembly, as this acts as a reference against which short read RNA sequencing data can be mapped [8].

* Correspondence: abhijeet.shah@uni-jena.de
[1]Institute of Ecology and Evolution, Friedrich Schiller University Jena, Dornburger Str. 159, 07743 Jena, Germany
[2]Department of Animal Behaviour, Bielefeld University, Morgenbreede 45, 33615 Bielefeld, Germany

Grasshoppers from the family Acrididae (Orthoptera, Caelifera) encompass a large number of species, including several economically relevant species, but have been poorly studied molecularly, partly because of their often very large genomes [9]. The 1Kite project has set out to sequence 1000 insect transcriptomes, including 43 species of Orthoptera, but only two Acridids (http://1kite.org/downloads/1KITE_species.txt). This is surprising because the Acridids are the largest family of Orthopterans, comprising around half of all Orthopteran species [10]. The only Acridid transcriptomes published so far are for the desert locust *Schistocerca gregaria* [11], the migratory locus *Locusta migratoria* [12], the stripe-winged grasshopper *Stenobothrus lineatus* [13] and the bow-winged grasshopper *Chorthippus biguttulus* [14].

The two locusts, *Schistocera* and *Locusta*, belong to the subfamilies Cyrtacanthacridinae and Oedipodinae respectively and have moderately large genomes [9]. Both species are of great interest due to their involvement in pest outbreaks and remarkable phenotypic plasticity. However, true phase-polymorphism and swarming behaviour is rather unusual, even among Acridids [15]. From an evolutionary perspective, the Acridids are remarkable for another phenomenon, the taxonomically widespread occurrence of an apparently balanced green-brown polymorphism in body coloration [16, 17].

The genera *Stenobothrus* and *Chorthippus* are representatives of the Gomphocerinae, a large subfamily of Acididae with particularly large genomes [9] and widespread green-brown polymorphisms, but no locust-like swarming behaviour. Of those two, only *Chorthippus biguttulus* is green-brown polymorphic.

Here, we present a de novo transcriptome assembly for the club-legged grasshopper *Gomphocerus sibiricus*, an alpine-dwelling species that is unusual for its striking sexual dimorphism in front leg morphology [18] (see Fig. 1) and is also characterized by the widespread occurrence of a balanced green-brown polymorphism [17]. The green-brown polymorphism has been found to be unrelated to rearing conditions [19] and recent evidence from the laboratory suggests that it follows a simple Mendelian mode of inheritance (unpublished data). Biliverdin is the dominant green pigment in Orthopterans [20] and it is likely that the biliverdin synthesis pathway plays an important role in the determination of body colour. However, the specific genes involved in producing the green-brown polymorphism are unknown in this species as well as in grasshoppers in general. Transcriptomic analysis thus offers a promising route for uncovering the genetic basis of this eco-evolutionarily relevant trait in club-legged grasshoppers as well as potentially in other species.

We assembled cDNA sequencing data from five *G. sibiricus* individuals including both males and females as well as specimens of the two colour morphs. In order to generate the most exhaustive assembly possible, we employed three different commonly used assemblers and combined the resulting assemblies into a single high-quality transcriptome assembly [21, 22]. We then compared our assembly to other published grasshopper assemblies and mined the transcripts for the presence of candidate genes involved in biliverdin pigmentation pathways. Additionally, we analysed the expression profile of the mitochondria and compared this to data from closely related species. Finally, having also found evidence for the presence of the parasitic microbe *Wolbachia* in our sample, we decided to assemble and analyse the *Wolbachia* strain present in this grasshopper species.

Fig. 1 The club-legged grasshopper *Gomphocerus sibiricus*, an alpine-dwelling species exhibiting prominent sexual dimorphism of the front leg. Photo credit: Holger Schielzeth

Results

Transcriptome assembly assessment and completeness metrics

We assembled a draft *G. sibiricus* transcriptome using three assemblers, followed by pooling of assemblies and removal of duplicates. The final draft assembly comprised 82,251 contigs, 21,347 of which contained open reading frames (ORFs). The TransRate transcriptome assembly metric was 0.325, which is similar to or better than most previously published orthopteran transcriptomes (Table 1). In order to further assess the quality of the transcriptome assembly, we constructed a kernel density plot of contig lengths versus the average depth of sequences mapping to those contigs (Additional file 1: Figure S1). Overall, the mapping rate using both read datasets was 96.72% (97.02% when using normalised reads, read mapping statistics for non-rRNA reads only

Table 1 TransRate assembly metrics and BUSCO completeness assessment for *Gomphocerus sibiricus* (this study) in comparison to the other two Acridid transcriptomes published (*Stenobothrus lineatus*, [13], *Chorthippus biguttulus*, [34]). Higher TransRate assembly scores indicate better quality assemblies. BUSCO completeness assessment was conducted using the insect ortholog database (orthoDB9)

TransRate metrics	*Gomphocerus sibirucus*	*Stenobothrus lineatus*	*Chorthippus bigullutus*
Number of contigs	82,251	57,778	67,733
Number of contigs with ORF	21,347	12,717	30,018
N50 of contig length	1357	1207	1246
Length of longest contig	43,026	22,561	34,437
Length of shortest contig	301	200	600
Proportion of read fragments mapped	0.88	0.70	0.51
Proportion of good read pairs mapping	0.82	0.63	0.43
TransRate assembly score	0.325	0.162	0.106
BUSCOs (Number of BUSCO units found)			
Complete BUSCOs	1405	1337	1489
Complete and Single-Copy BUSCOs	1093	1244	1323
Complete and duplicated BUSCOs	312	93	166
Fragmented BUSCOs	137	142	99
Missing BUSCOs	116	179	70
Total BUSCOs searched	1658	1658	1658

are available in Additional file 7: Table S3) with BWA under default settings. The mean (median) contig length was 1057 bp (718 bp) and the mean (median) coverage was 52.6x (11x), indicating that we were able to assemble relatively complete contigs.

Transcriptome annotation and GO classification

Over 87% of the contigs were annotated using InterProScan 5, BLASTx (non-redundant protein database) and the *dammit!* annotation pipeline. The InterProScan analysis resulted in 1,588,733 matches from 72,132 annotated sequences, with 337,258 assigned membrane-bound protein signatures. Furthermore, 109,437 GO terms were assigned to 39,794 transcripts. A BLASTX search of transcripts was run against the non-redundant (nr) protein database which yielded 12,436,946 hits from 31,527 transcripts. Around 78% (24,568) of these transcripts were assigned to insects, and the next largest fraction (3.8%) was assigned to the Arachnida (Additional file 2: Figure S2). The *dammit!* annotation pipeline provided us with 363,218 annotations from 44,488 transcripts. Of these, 41,535 transcripts were annotated by both the InterProScan 5 and *dammit!* Pipelines (Additional file 5: Table S1).

The annotated draft transcriptome was classified into three main categories of GO components: cellular components, molecular function and biological processes (Additional file 3: Figure S3). Of these, 3617 (9.1%) were classified as cellular component, 26,974 (67.8%) as molecular function and 9203 (23.2%) as biological process. Furthermore, GO terms associated with accumulation of pigment and haeme biosynthesis (GO:0043473 pigmentation; GO:0006783, GO:0006784 haeme; GO:0048034 haeme-O and haeme oxygenase decyclizing activity) were found amongst these annotations. Additionally, we found several relevant InterPro terms: IPR015118 (5-aminolevulinate synthase presequence), IPR010961 (tetrapyrrole biosynthesis of 5-aminovulenic acid), IPR002051 (haeme oxygenase) and KEGG reference: 00860 + 1.14.14.18 (biliverdin producing haeme oxygenase) in our annotations. This provides a starting point for investigating differential gene expression in relation to the green-brown colour polymorphism in future studies.

SNP calling and estimation of minor all frequencies (MAF)

We found 76,221 SNPs in 84,00 contigs with mean of 9.07 variants per contig. A contour plot shows that minor allele frequencies (MAFs) peak between 0.12 and 0.2, while rarer variants could only be detected with increasing depth of sequencing coverage (Fig. 2).

Scanning for expressed mitochondrial genes in assembled transcripts

G. sibiricus mitochondrial sequences were also found in the assembled transcripts, and their coverage depth was estimated. An alignment of the assembled draft transcripts against the published mitochondrial genome [23] shows that we could recover large

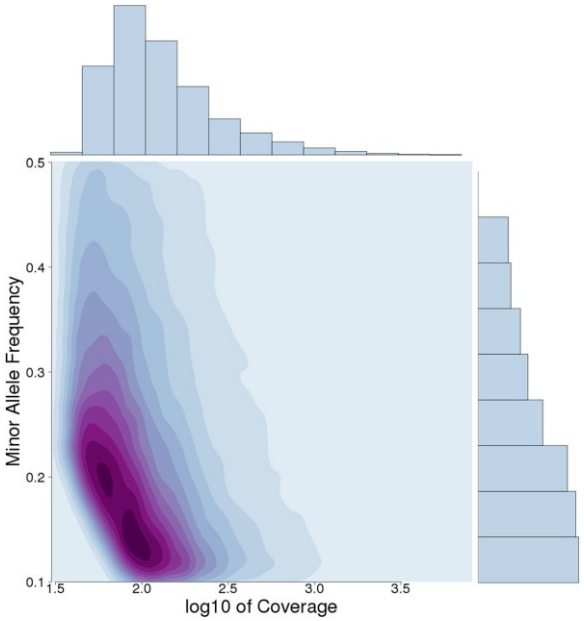


Fig. 2 A two-dimensional kernel density plot showing minor allele frequency (MAF) plotted against log10 sequence coverage. The dark purple regions indicate higher densities, whereas the light blue regions indicate lower densities. Marginal histograms show frequency distributions of minor allele frequency (top axis) and coverage (right axis)

multi-gene segments. The published mitochondrial genome of *G. sibiricus* (refseq: NC_021103.1) consists of a standard set of 13 protein coding genes and two ribosomal RNA genes (16 s and 12 s ribosomal sub-units). As expected, our contigs mapped to the reference across the full mitochondrial genome with particularly high coverage over the 16 s and 12 s ribosomal RNA genes (Fig. 3).

The distance matrix of the mitochondrial genome assembled from this study and four closely related taxa provides an overview of the amount of sequence divergence within the genus *Gomphocerus*. As expected, the lowest divergence was found between our European *G. sibiricus* population and the Asian population of the same species (divergence 1.6%, Additional file 6: Table S2), while there was slightly more divergence with the Asian congeneric species *G. licenti* and *G. tibetanus*, and the largest divergence was found in the comparison with *G. rufus* (which is usually placed within the genus *Gomphocerippus*). This analysis also uncovered two unexpected patterns. First, *G. sibiricus* was found to be less divergent from *licenti* than from *tibetanus*, even though the latter is sometimes considered a subspecies of *G. sibiricus* [10]. Second divergence estimates with *licenti* / *tibetanus* were lower than with the published *G. sibiricus* sequence from Central Asia.

Detection of Wolbachia PeI wPip strain sequences

We detected the endosymbiont *Wolbachia* in our RNA sequencing data. For strain determination, we mapped our *Wolbachia* assembly to all the *Wolbachia* genomes available on GenBank (8280 assemblies, retrieved 30/11/2017) using *BWA* aligner. Our *Wolbachia* strain showed the best mapping to strain wPip, which was originally described from the *Culex quinquefasciatus* Pel genome [24]. Mean read coverage was quantified using a sliding window of 250 bp (Fig. 4). The coverage of *G. sibiricus* transcripts suggests ample uniform coverage, with two major peaks close to genome positions 11,360,000 and 12,360,000, which correspond to 16 s and 23 s ribosomal RNA respectively. Additionally, a table containing the transcripts which have the top BLASTX hits to *Wolbachia* are provided in table in an Additional file 8.

Discussion

We here present a high quality transcriptome assembly for *G. sibiricus*, a species of grasshopper with a large genome that is of particular evolutionary interest

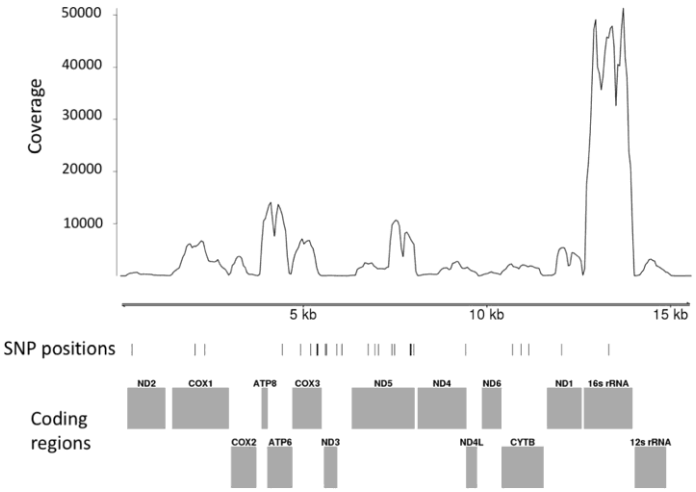


Fig. 3 Transcripts from *Gomphocerus sibiricus* mapping to the reference mitochondrial genome assembly. The short vertical lines indicate positions of putative SNPs. The grey shaded blocks indicate protein and rRNA coding regions

because it features a green-brown polymorphism that is shared with many other Orthopterans and seems to be maintained in natural populations by balancing selection [17]. Assembly quality statistics from BUSCO and TransRate demonstrate that the assembly is of similar quality to the only two available assemblies for the subfamily Gompocerinenae, which are also characterized by large genome sizes (the estimated genome size of *G. sibiricus* is approximately 8.75 Gb) [25]. Our draft transcriptome provides a resource for future studies of differential gene expression, which should ultimately help to improve our understanding of the genetic basis of colour morph determination. Furthermore, our mitochondrial and *Wolbachia* sequence assemblies provide material for phylogenetic and comparative analyses.

Previously, a comparison of Gomphocerinae mitochondrial genomes [23] suggested that *G. sibiricus* is more closely related to *G. licenti* than to *G. tibetanus*, although *tibietanus* is sometimes listed as a subspecies of *sibiricus*, while *licenti* is considered a separate species

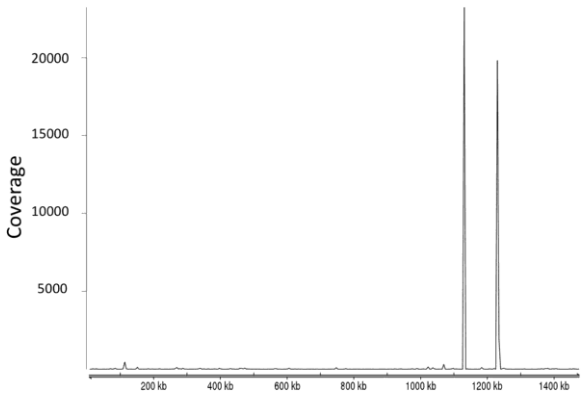


Fig. 4 Coverage of the *Wolbachia* genome estimated by mapping *G. sibiricus* transcripts against the pel wPi strain of *Wolbachia*. The major peaks around positons 1,136,000 bp and 1,236,0000 bp correspond to the 16 s and 23 s rRNA coding regions

[10]. Our analysis with an independent sample of *sibiricus* confirms the results of Zhang et al. [23], as we also found less divergence from *licenti* than from *tibetanus*. Despite the greater geographical distance, however, we find our sample to be closer to *licenti* / *tibetanus* than the previously published *G. sibiricus* mitochondrial sequence from China. This may be related to the quality of the assembly (erroneous assemblies will tend to reduce similarity) or it could potentially reflect a yet unexplored and surprising colonization history from Central Asia to the mountains in Europe.

We also found the first evidence that *G. sibiricus* is infected by *Wolbachia*, a microbial parasite that is known to manipulate host reproductive biology and can cause horizontal gene transfer [26, 27]. Our finding is not entirely unexpected given that many insect species are known to be infected by *Wolbachia* [28]. Moreover, *Wolbachia* infections have recently been reported in other grasshopper species including *Chorthippus parallelus* [29, 30] and *Podisma saporensis* [31]. *Wolbachia* causes cytoplasmic incompatibility in *C. parallelus* [30] and is suspected to play a leading role in causing hybrid dysfunction in *P. saporensis* [31]. Our draft transcriptome assembly therefore provides a basis for future studies investigating the evolutionary and ecological significance of *Wolbachia* infections in *G. sibiricus*.

As we were able to annotate the majority of our assembled transcripts, our assembly provides a useful tool for investigating the genetic basis of complex traits such as pigmentation. Our GO annotations allowed us to identify putative candidate genes for investigating the mechanism of colour morph determination. We demonstrated the presence of multiple transcripts that are putatively involved in pigmentation, more specifically, in the porphyrin and chlorophyll metabolism pathway (KEGG ko00860) and in haeme oxygenase 2 (biliverdin-producing, HMOX2) [32]. Furthermore, we found indirect evidence for precursors and other metabolites involved in pigmentation pathways, possibly reflecting the lack of available metabolic and biochemical resources and the significant phylogenetic distance of *G. sibiricus* from well-annotated model organisms. Importantly, the GO and InterPro evidence for transcripts involved in haeme and haeme-O complex metabolism (GO:0006783, GO:0006784, GO:0048034, IPR015188, IPR010961) suggests that components for a plausible mechanism to metabolize green pigments from plant sources exist in this species.

Our transcriptome assembly of *G. sibiricus* combines the best assembled transcripts from three different de novo transcriptome assemblers (with multiple k-mer assemblies) to detect, capture and assemble transcripts. Recently, Smith-Unna et al. [33] investigated de novo transcriptome assembly quality from 155 previously published transcriptome assemblies. They found that assemblies generated by individual assemblers yielded relatively low TransRate scores, but when multiple assemblies were combined, reduced and filtered, they yielded reasonably representative collections of sequenced read fragments. According to this particular quality metric, our draft transcriptome assembly for *G. sibiricus* lies in the approximate upper 70th percentile of the surveyed published transcriptomes, suggesting that the quality of our assembly is well above average. In our case, the combination of different assemblers was key to effective transcriptome assembly.

Conclusions

In conclusion, we generated a high quality transcriptome assembly for *G. sibiricus*, a species with a large genome and limited available molecular resources. Not only will our study provide a solid foundation for studying the genetic basis of green-brown colour dimorphism in *G. sibiricus*, but we also generated resources that should facilitate future studies of mitochondrial genome evolution and *Wolbachia* infections of Orthopterans.

Methods

Sample collection

Individuals of *Gomphocerus sibiricus* were collected from a field site at 1800–2000 m near Sierre (Valais, Switzerland) in 2013. Five individuals of their laboratory-reared offspring were selected for RNA sequencing. This included one imago brown female, one imago green female, one imago brown male, one imago green male and one last-instar green female. Sexes, colour morphs and developmental stages were mixed in order to achieve a sufficient representation of the species' transcriptome.

RNA extraction, cDNA library preparation and spike-in normalisation (step 1)

Total RNA was extracted using an innoPrep RNA kit (Analytik Jena, Jena, Germany), followed by quality control and quantification using the RNA 6000 Nano LabChip kit with the Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). The samples were pooled before cDNA library for transcriptome sequencing was constructed. In order to maximize sequencing efforts, the poly-(A)-containing mRNA was purified from total RNA using poly-(T) oligo-attached magnetic beads. This was followed by cDNA library construction, which was prepared and sequenced by the Center for Biotechnology at Bielefeld University on the Illumina MiSeq platform (San Diego, California, USA) with a maximum read length of 300. Next, an aliquot of the cDNA library was sent to a commercial facility for standard ERCC spike-in normalization, followed by an additional round of sequencing using the same protocol, to ensure that sequencing

efforts were not only concentrated on highly abundant sequences [34]. The normalised library was used for transcriptome assembly, while the non-normalised library was used for read mapping and transcript confirmation.

Pre-processing and sequence quality control and k-mer filtering (step 2)

All pair-end reads were processed and trimmed to a maximum length of 300 bp to remove low-quality bases. The resulting reads were assessed for quality using FastQC (version 0.11.15). Sequence quality and adapter trimming was performed using the trimmomatic tool (version 0.36) [35], with a four base sliding window. Bases below a phred quality score of 15 were removed as were reads with a sequence length below 150 using the setting '2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:150'. Furthermore, the reads were in silico normalized using Khmer (version 2.0) following the authors' recommended protocol for trimming reads with variable coverage (recipe 7) [36]. Additional artefact filtering was carried out using the FASTX toolkit (version 0.0.14). This resulted in a final read dataset of 5,070,036 paired-end reads with a median read length of 249 bp for the normalised read dataset.

Transcriptome assemblies (step 3)

We assembled the *G. sibiricus* transcriptome using the de novo transcriptome assembly packages SOAPdenovo-trans [37], Trinity [38] and Oases-Velvet [39]. An overview of this process is given in Fig. 5. First, we used SOAPdenovo-trans (version 1.03) to build 54 de novo assemblies with k-mer values of 21 to 127 (in steps of 2) with an average insert size of 475 bp, using -F and -L 300 command parameters to set filling the gaps in scaffolds and shortest contig length for scaffolding to 300 bp. Second, we used Trinity (release v2.2.0) with the default settings, the minimum transcript length set to 300 bp, and with *in-silico* read normalization with default settings. Third, Oases-Velvet (version 0.2.09) was used to build eighteen de novo assemblies with k-mer values of 21 to 55 (in steps of 2) with an insert length of 475 bp and the minimum transcript length set to 300 bp. Assemblies with higher k-mer values were not possible due to memory constraints.

Multi-assembly merging and duplicate contig reduction (step 4)

All of the assemblies were collated into a single file and followed by removal of duplicate and redundant contigs. CD-HIT (cd-hit-est, version 4.6) [40], a greedy incremental clustering algorithm, was used to remove duplicate and highly similar sequences (sequence similarity greater than 90% identity) to generate representative sequences of all of the assembled transcripts from all of the assemblies. This approach allowed us to capture as many unique transcripts as possible from three different de novo transcriptome assemblers.

Transcript mapping and quantification (step 5)

The quality filtered reads were mapped to our de novo transcriptome assembly and the number of mapping reads per transcript was quantified using

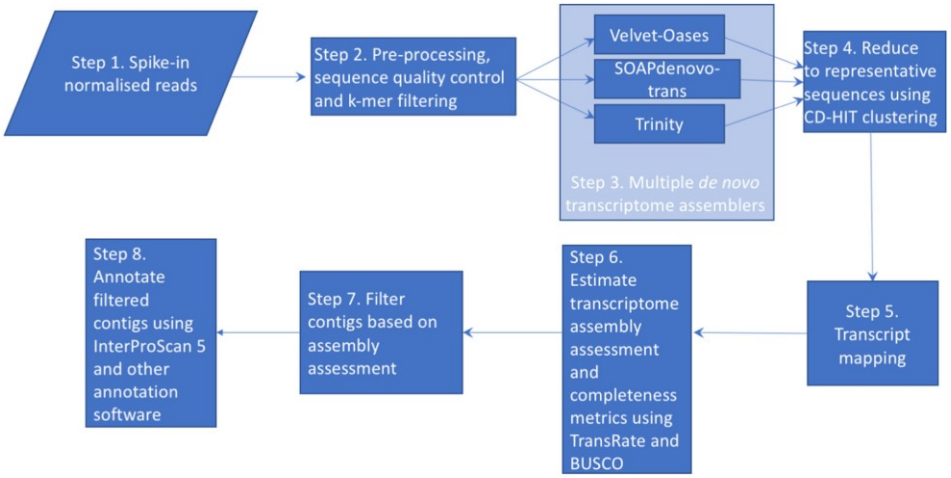


Fig. 5 Flow chart showing a summary of the de novo assembly procedure and downstream analyses

BWA (version 0.712-r1039) [41]. Per contig and per base coverage was calculated using BBMap [42] and SAMtools (version 1.4) [43].

Assessment of assembled transcriptome and contig filtering (step 6 and 7)

We assessed the quality of the de novo assembly using the TransRate package (version 1.0.3) [33]. This assesses transcriptome assembly accuracy using read and contig data and estimates individual contig and overall assembly quality. Contigs that scored low on the contig assessment criteria (contig score components: $S(C_{seg})$ and $S(C_{cov})$, with default thresholds as suggested by TransRate), were filtered out. We then compared our empirical assembly metrics to the published transcriptomes of two other Gomphocerine species for which transcriptome assemblies have been published (*Stenobothrus lineatus* [13] and *Chorthippus biguttulus* [44]). Assembly completeness was assessed with Benchmarking Universal Single-Copy Orthologs (BUSCO, version 2.0) [45]. BUSCO defines a set of core genes for a given group or lineage and uses these genes as proxy for minimum completeness assuming that these genes should encode a large set of core genes. For this analysis, we used the genome and transcriptome completeness assessment tool in transcriptome assessment mode with the insect lineage database (insecta_orthoDB9, created 13/02/2016).

Annotation and sequence analysis (step 8)

After assembly and assembly assessment, we annotated and analysed the assembly. InterProScan 5 is a widely used sequence analysis framework to search for various analytical signatures from different databases including the InterPro protein database. The draft transcript sequence signatures were scanned against InterPro's signatures (version 5.22) [46]. TMHMM, SignalP, TIGRFAMs, Prosite, Panther, PFam, PIRSF, CDD, COILS, and Gene3D applications were selected to run with the transcripts using the InterScanPro 5 framework. Additional annotation was generated using the *dammit!* annotation pipeline (https://github.com/dib-lab/dammit/). This pipeline was set to use the Pfam, Rfam, OrthoDB, uniref90 and BUSCO arthropoda databases in the '--full' mode. All assembled contigs were also searched against the BLASTx non-redundant protein database (Nr) and results with expect values (e-values) below 10^{-6} were reported. To visualize the BLAST profile of the assembly, we selected the best matches (based on best bit-scores) from every transcript and plotted the counts of taxonomic classes found.

SNP calling and estimation of minor allele frequencies

We detected SNPs from the transcriptome by mapping the quality-filtered reads to the final set of assembled transcripts using the *BWA* aligner (version 0.712-r1039) [47] (using default settings) followed by variant detection with VarScan (version 2.4.2) [48] with minimum coverage set to 30, a minimum expected variant frequency set to 0.1, and *p*-value threshold of 0.01. In order to visualize our variants with respect to sequencing depth, we estimated the two-dimensional kernel density of the minor allele frequency and plotted this as a contour plot.

Searches for candidate green-brown pigmentation genes

Gene ontology terms (GO terms) associated with pigmentation (GO:0043473) and haeme-metabolism (GO:0004392, GO:0006784 and GO:0048034) were selected as possible candidates for genes associated with the observed green-brown dimorphism. Furthermore, InterPro and KEGG terms (IPR010961, IPR015118, IPR002051, KEGG00860 + 1.14.14.18) associated with haeme oxygenase (the pathway leading to biliverdin production) were also added to the annotation search. The presence of these GO terms illustrates that putative candidates for pigmentation are represented in our assembly.

Analysis of the mitochondrial genome

We also assembled the mitochondrial transcriptome of *G. sibiricus*. Reads were mapped to a the reference mitochondrial genome from a specimen of the same species from China (NCBI Reference Sequence NC_021103.1) [23]. This was done using *BWA* with default settings, and coverage was estimated using BBMap [42] and SAMtools (version 1.4) [43]. Furthermore, the assembled transcripts were BLASTed against the reference mitochondrial genome to screen for potential chimeric components in the assembled transcripts. Variant calling was again performed using VarScan (version 2.4.2) [48] and all of the results were plotted and annotated using the Gviz package [49] on Bioconductor [50].

Finally, we aligned and estimated the sequence distance of four closely related Gomphocerinae mitogenomes, namely *G. rufus* (RefSeq: NC_014349), *G. licenti* (RefSeq: NC_013847), *G. tibetanus* (RefSeq: NC_015478), *G. sibiricus* from the Tianshan Mountians, Xinjiang, China (RefSeq: NC_021103) and our data from *G. sibiricus* from the European Alps. The multiple sequence alignment was estimated using *MAFFT* [51] with the default setting in 'auto' mode. The program *dnadist* from the *PHYLIP* package (version 3.696) [52] was used with default settings (F84 substitution model, transition/transversion ratio 2.0, using empirical base frequencies) to estimate pairwise sequence distances. The program *dnaml* from the PHYLIP package [52] was used to estimate the best fitting phylogenetic tree with default settings (transition/transversion

ratio 2.0, constant rate variation among sites), with randomized input order of sequences, and *G. rufus* set as the outgroup and plotted (see Additional file 4: Figure S4).

Additional files

Additional file 1: Figure S1. Log of average fold coverage versus log of contig length. (DOCX 58 kb)

Additional file 2: Figure S2. Top 20 taxa classes reported by BLASTX. (DOCX 50 kb)

Additional file 3: Figure S3. Classifications of GO terms of contigs. (DOCX 251 kb)

Additional file 4: Figure S4. A phylogeny based on mitochondrial sequences of four Gomphocerine grasshopper species. (JPG 35 kb)

Additional file 5: Table S1. A Summary table of the annotation of the contigs using the *dammit!* Pipeline. (DOCX 13 kb)

Additional file 6: Table S2. Sequence divergence matrix from four Gomphocerine grasshopper species. (DOCX 13 kb)

Additional file 7: Table S3. Read mapping statistics for non-rRNA reads. (DOCX 12 kb)

Additional file 8: Table of top Wolbachia BLASTX hits from the assembly. (CSV 104 kb)

Abbreviations
BLAST: Basic local alignment tool; bp: base pair; BUSCO: Benchmarking Universal Single-Copy Orthologs; ERCC: External RNA Control Consortium; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; MAF: Minor Allele Frequency; ORF: Open Reading Frames

Acknowledgements
We are grateful to Amy R. Backhouse for carrying out the sample preparations and RNA extractions and Anika Winkler for further lab assistance. We would also like to thank Drs. Emily Humble and Luke Eberhart-Philips for assisting us in plotting data. We also thank The Center for Biotechnology (CeBiTec) at Bielefeld University to provide compute cluster resources to facilitate this endeavour.

Funding
This study was supported by an Emmy-Noether fellowship by the German Research Foundation (DFG; SCHI 1188/1–1). The funding body played no role in the design of the study, collection of samples, analysis and interpretation of data, and writing the manuscript.

Availability of data and materials
The draft transcriptome and read data are deposited in the Transcriptome Shotgun Assembly Sequence database and in the Short Read Archive respectively, under BioProject ID:PRJNA525981 with the accession id: GHKV00000000.

Authors' contributions
AS, HS and JIH, conceived and designed the project. AS executed the bioinformatic pipelines and performed the data analysis. AS, JIH and HS drafted the manuscript. All authors contributed to revision and approved the final version.

Ethics approval and consent to participate
The authors declare that the experiments comply with the current law of the country in which they had been performed. All sampling was carried out with non-protected species in non-protected areas.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 30 May 2018 Accepted: 30 April 2019
Published online: 14 May 2019

References
1. Lee YW, Gould BA, Stinchcombe JR. Identifying the genes underlying quantitative traits: a rationale for the QTN programme. Aob Plants. 2014;6:plu004.
2. Schielzeth H, Husby A. Challenges and prospects in genome-wide QTL mapping of standing genetic variation in natural populations. Ann N Y Acad Sci. 2014;1320:35–57.
3. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010;20(9):1165–73.
4. Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG, et al. The axolotl genome and the evolution of key tissue formation regulators. Nature. 2018;554(7690):50–5.
5. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–82.
6. Parmigiani G, Garrett ES, Irizarry RA. Zeger SL (eds.): the analysis of gene expression data : methods and software. New York: Springer; 2003.
7. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
8. Zhang R, Calixto Cristiane PG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, Spensley M, Entizne JC, Lewandowska D, ten Have S, et al. A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res. 2017;45(9):5061–73.
9. Gregory TR. Animal Genome Size Database; 2018.
10. Cigliano MM, Braun H, Eades DC, Otte D. Orthoptera Species File (Version 5.0/5.0). In. 2018.
11. Badisco L, Huybrechts J, Simonet G, Verlinden H, Marchal E, Huybrechts R, Schoofs L, De Loof A, Vanden Broeck J. Transcriptome analysis of the desert locust central nervous system: production and annotation of a Schistocerca gregaria EST database. PLoS One. 2011;6(3):e17274.
12. Wang XH, Fang XD, Yang PC, Jiang XT, Jiang F, Zhao DJ, Li BL, Cui F, Wei JN, Ma CA, et al. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 2014;5:2957.
13. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. Phylogenomics resolves the timing and pattern of insect evolution. Science. 2014;346(6210):763–7.
14. Berdan EL, Finck J, Johnston PR, Waurick I, Mazzoni CJ, Mayer F. Transcriptome profiling of ontogeny in the acridid grasshopper Chorthippus biguttulus. PLoS One. 2017;12(5):e0177367.
15. Song H. Density-dependent phase polyphenism in nonmodel locusts: a minireview. Psyche. 2011;2011 741769.
16. Bellmann H, Luquet CH. Guide des sauterelles, grillons et criquets d'Europe occidentale. Paris: Delachaux et Niestlé; 2009.
17. Dieker P, Beckmann L, Teckentrup J, Schielzeth H. Spatial analyses of two colour polymorphisms in an alpine grasshopper reveal a role of small-scale heterogeneity. Ecol Evol. 2018. https://doi.org/10.1002/ece1003.4156.
18. Valverde JP, Eggert H, Kurtz J, Schielzeth H. Condition-dependence and sexual ornamentation: effects of immune challenges on a highly sexually dimorphic grasshopper. Insect Sci. 2017. https://doi.org/10.1111/1744-7917.12448.
19. Valverde JP, Schielzeth H. What triggers colour change? Background colour and temperature effects on the development of an alpine grasshopper. BMC Evol Biol. 2015;15:168.
20. Fuzeau-Braesch S. Pigments and color changes. Annu Rev Entomol. 1972;17:403–24.
21. Cerveau N, Jackson DJ. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. BMC Bioinformatics. 2016;17(1):525.
22. Haznedaroglu BZ, Reeves D, Rismani-Yazdi H, Peccia J. Optimization of de novo transcriptome assembly from high-throughput short read sequencing data improves functional annotation for non-model organisms. BMC Bioinformatics. 2012;13(1):170.
23. Zhang H-L, Zhao L, Zheng Z-M, Huang Y. Complete mitochondrial genome of Gomphocerus sibiricus (Orthoptera: Acrididae) and comparative analysis in four Gomphocerinae mitogenomes. Zool Sci. 2013;30(3):192–204.

24. Klasson L, Walker T, Sebaihia M, Sanders MJ, Quail MA, Lord A, Sanders S, Earl J, O'Neill SL, Thomson N, et al. Genome evolution of Wolbachia strain wPip from the Culex pipiens group. Mol Biol Evol. 2008;25(9):1877–87.
25. Gosalvez J, López-Fernandez C, Esponda P. Variability of the DNA content in five Orthopteran species. Caryologia. 1980;33(2):275–81.
26. Ahmed MZ, Breinholt JW, Kawahara AY: Evidence for common horizontal transmission of Wolbachia among butterflies and moths. BMC Evol Biol 2016, 16(1):118–118.
27. Werren JH, Baldo L, Clark ME. Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol. 2008;6(10):741–51.
28. Hilgenboecker K, Hammerstein P, Schlattmann P, Telschow A, Werren JH. How many species are infected with Wolbachia? – a statistical analysis of current data. FEMS Microbiol Lett. 2008;281(2):215–20.
29. Zabal-Aguirre M, Arroyo F, Bella JL. Distribution of Wolbachia infection in Chorthippus parallelus populations within and beyond a Pyrenean hybrid zone. Heredity. 2009;104:174.
30. Zabal-Aguirre M, Arroyo F, García-Hurtado J, Torre J, Hewitt GM, Bella JL. Wolbachia effects in natural populations of Chorthippus parallelus from the Pyrenean hybrid zone. J Evol Biol. 2014;27(6):1136–48.
31. Bugrov AG, Ilinsky YY, Strunov A, Zhukova M, Kiseleva E, Si A, Tatsuta H. First evidence of Wolbachia infection in populations of grasshopper Podisma sapporensis (Orthoptera: Acrididae). Entomological Science. 2016;19(3):296–300.
32. Ishikawa K, Takeuchi N, Takahashi S, Matera KM, Sato M, Shibahara S, Rousseau DL, Ikeda-Saito M, Yoshida T. Heme Oxygenase-2: properties of the heme complex of the purified tryptic fragment of recombinant human heme oxygenase-2. J Biol Chem. 1995;270(11):6345–50.
33. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. 2016;26(8):1134–44.
34. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011; 21(9):1543–51.
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
36. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A, Constantinides B, Edvenson G, Fay S, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research. 2015;4:900.
37. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, et al. SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30(12):1660–6.
38. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.
39. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086–92.
40. Li WZ, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
41. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXive preprint 2013:available at https://arxiv.org/abs/1303.3997.
42. Bushnell B. BBMap short read aligner. In: University of California, Berkeley, California; 2016.
43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Genome project data processing S: the sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
44. Berdan EL, Mazzoni CJ, Waurick I, Roehr JT, Mayer F. A population genomic scan in Chorthippus grasshoppers unveils previously unknown phenotypic divergence. Mol Ecol. 2015;24(15):3918–30.
45. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
46. Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40.
47. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–60.
48. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17): 2283–5.
49. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. In: Mathé E, Davis S, editors. Statistical genomics: methods and protocols. New York, NY: Springer New York; 2016. p. 335–51.
50. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, et al. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015;12(2):115–21.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80.
52. Felsenstein J. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat. 2008;171(6):713–25.

Manuscript III

Title: Comparative analysis of genomic repeat content in gomphocerine grasshoppers reveals expansion of satellite DNA and helitrons in species with unusually large genomes

Authors: Abhijeet Shah, Joseph Hoffman and Holger Schielzeth

Abstract

Eukaryotic organisms vary widely in genome size and much of this variation can be explained by differences in the abundance of repetitive elements. However, the phylogenetic distributions and turnover rates of repetitive elements are largely unknown, particularly for species with large genomes. We therefore used *de novo* repeat identification based on low coverage whole-genome sequencing to characterize the repeatomes of six species of gomphocerine grasshoppers, an insect clade characterised by unusually large and variable genome sizes. Genome sizes of the six species ranged from 8.4 to 14.0 pg DNA per haploid genome and thus include the second largest insect genome documented so far (with the largest being another acridid grasshopper). Estimated repeat content ranged from 79 to 96% and was strongly correlated with genome size. Averaged over species, these grasshopper repeatomes comprised significant amounts of DNA transposons (24%), LINE elements (21%), helitrons (13%), LTR retrotransposons (12%) and satellite DNA (8.5%). The contribution of satellite DNA was particularly variable (ranging from <1% to 33%) as was the contribution of helitrons (ranging from 7 to 20%). The age distribution of divergence within clusters was unimodal with peaks around 4–6%. The phylogenetic distribution of repetitive elements was suggestive of an expansion of satellite DNA in the lineages leading to the two species with the largest genomes. Although speculative at this stage, we suggest that the expansion of satellite DNA could be secondary and might possibly have been favoured by selection as a means of stabilising greatly expanded genomes.

Introduction

Large fractions of eukaryotic genomes consist of repetitive elements, which vary considerably in their abundance across species (Charlesworth *et al.*, 1994; Lynch & Conery, 2003). The repetitive fraction of the genome, known as the *repeatome*, correlates with genome size both within and among species (Lynch, 2007) and therefore likely plays a major role in genome size evolution (Charlesworth *et al.*, 1994; Talla *et al.*, 2017). Some repeats, such as transposable elements, spread as selfish elements that do not benefit the host organism (Doolittle & Sapienza, 1980; Orgel & Crick, 1980). However, repeats are also known to assume functional roles (Shapiro & von Sternberg, 2005), such as centromeric satellite DNA, which is necessary for appropriate chromosome pairing during cell division (Hartl, 2000; Plohl *et al.*, 2008). Repeat elements have also been associated with genetic innovation and speciation (Ellegren *et al.*, 2012; Feliciello *et al.*, 2014; Maumus *et al.*, 2015), rendering repeatome analysis relevant to understanding the origin and maintenance of biodiversity in general.

A small number of clades have evolved genome size gigantism including some gymnosperms, amphibians, crustaceans, lungfish, sharks, velvet worms, flatworms and grasshoppers (Gregory, 2018). Despite these independent origins of extreme genome size expansions, most species have rather compact genomes (Gregory, 2018). Overall, genome size does not appear to be related to organismal complexity, a disparity that is known as the C-value enigma because genome size is typically quantified by the C value (the molecular weight of a haploid genome, (Gregory, 2005)). Instead, certain factors or circumstances may have allowed genome sizes to increase in some groups but not in others, although these conditions are in general poorly understood. A comparative analysis of the repeatomes of species with large genomes may therefore shed light on the C-value enigma and contribute towards an improved understanding of genome size expansions.

A desirable approach would be to conduct a comparative analysis of assembled and annotated genomes in which specific repetitive elements can be clearly identified. However, it is precisely the repeat content that has hindered the assembly of reference genomes for species with large genomes (Plohl *et al.*, 2012; Ruiz-Ruano *et al.*, 2017). The largest genomes published so far are draft genomes of the migratory locust *Locusta migratoria* (6.38 Gb

(Wang *et al.*, 2014)), Norway spruce *Picea abies* (19.6 Gb, (Nystedt *et al.*, 2013)) and Mexican axolotl *Ambystoma mexicanum* (32.39 Gb, (Nowoshilow *et al.*, 2018)). The case of the migratory locust illustrates the difficulty of assembling large and repetitive genome sequences, as the current assembly is fragmented into more than 550,000 scaffolds with an N50 of 322 kb, despite only 12 chromosomes contributing to the species' large genome size (Wang *et al.*, 2014). The difficulty of assembling repetitive regions in particular has hampered progress in the analysis of repetitive elements in such species.

Recent comparative studies on genome sizes in insect have focused on the entire group at large and included the migratory locust as the only orthopteran with the largest genome in the sample (Petersen *et al.*, 2019; Wu & Lu, 2019). Here, we use a comparative approach to study repeat content in a group of grasshoppers that has genome sizes exceeding that of the migratory locust. We chose to study grasshoppers of the subfamily Gomphocerine (Orthoptera, suborder Caelifera, family Acrididae) because they have highly variable genome sizes, both across and in some cases within species (Gregory, 2018; Jetybayev *et al.*, 2018; Schielzeth *et al.*, 2014). This clade hosts the largest genomes among all insects and, even across all organisms, it represents one of only a small set of clades with extremely large genomes (Gregory, 2018). Although this makes genome assembly challenging for orthopterans, it offers an outstanding opportunity for a comparative analysis of the repeatome.

The short-horned grasshoppers (Caelifera) have a rather conserved basic karyotype with 9 or 12 chromosome pairs (John & Hewitt, 1966), so that genome size variation across species are largely due to differences in the sizes rather than the numbers of chromosomes. At the same time, grasshoppers often vary intra-specifically in chromosome number (Palestis *et al.*, 2004). Supernumerary chromosomes (B chromosomes) and chromosomal segments consist mostly of heterochromatin, which is rich in repeats, especially satellite DNA (Ruiz-Ruano *et al.*, 2017; Ruiz-Ruano *et al.*, 2018). Consequently, grasshoppers show stark contrasts between phylogenetically conserved karyotypes, substantial variation in chromosome size, and facultative variation in dispensable DNA segments. The frequent presence of large pieces of additional DNA also suggests that mechanisms of genome size control are rather weak and/or that tolerance to increases in genome size is high.

We used whole-genome shotgun sequencing to characterise the repeatomes of six species of gomphocerine grasshoppers (Figure 1). With low-coverage sequencing it is unlikely that sequences with single copies in the genome will be represented multiple times in the data. Repeated sequences with hundreds or thousands of copies, however, are represented by multiple reads even when sequencing coverage is low. Comparative *de novo* assembly of low coverage sequences therefore facilitates the assembly of the repetitive fraction of the genome and thus provides insights into the types and distributions of repetitive DNA. We used a multi-stage analytical pipeline incorporating graph-based *de novo* clustering of repeat elements (Figure S1) building on the software packages RepeatExplorer (Novák *et al.*, 2013) and dnaPipeTE (Goubert *et al.*, 2015b) as well as RepeatMasker (Smit, 2015) and RepBase (Bao *et al.*, 2015) for annotation.

We recently analysed the repeat content of one species of gomphocerine grasshopper, the club-legged grasshopper *Gomphocerus sibiricus* (Shah *et al.*, 2016). The distribution of repeat types across read clusters of transposable element copies differed markedly from other published distributions (e.g. (da Silva *et al.*, 2018; Lower *et al.*, 2017; Piednoel *et al.*, 2012)) in that this species shows a large dominance of one particular cluster annotated as satellite DNA. The existence of one predominant class of repeats argues for a recent expansion of this type of repeat sequence in the focal genome, because with an ancient expansion, we would have expected the repeat sequences to have diverged by mutation, which would result in them assembling into multiple clusters rather than into a single cluster. One motivation for the current analysis was therefore to determine whether satellite DNA repeats also appear at high frequency in the genomes of related grasshopper species.

We tested the prediction that grasshopper repeatomes show a strong phylogenetic signal, being more similar in closely related species, while also searching for particular repeat classes showing signs of expansion or reduction in specific lineages. Furthermore, we aimed to evaluate if the unusual pattern of striking dominance of satellite DNA in the genome of *G. sibiricus* is species-specific or represents a more general characteristic of gomphocerine grasshoppers. By analysing a suite of species that vary substantially in their genome sizes, we aimed to test for a relationship across species between genome size and repeat content.

Finally, by analysing sequence divergence within clusters, we attempted to evaluate the relative ages of expansions of particular repeat classes.

Results

We combined low coverage short-read sequencing with graph-based clustering to characterize the relative abundances of the most common repeats across six species of gomphocerine grasshoppers (Figure 1). For brevity, and because genus assignment has recently been in flux, we hereafter refer to each taxon only by its species name (*parallelus*, *variegatus*, *biguttulus*, *rufus*, *sibiricus* and *scalaris*, respectively). Genome size was determined by flow cytometry using the house cricket *Acheta domesticus* as a size standard (2.1 pg DNA per haploid genome). We found that genome size varied across species by a factor of 1.7, with *scalaris* having the largest genome (approx. 14.0 pg) and *biguttulus* the smallest (approx. 8.4 pg, Figure S2). Sequencing of 12 individuals, comprising one individual of each sex from six different species, resulted in a total of approximately 311 million reads, which after quality filtering was reduced to approximately 300 million reads (20.4–43.0 million reads per sample) totalling 34.1 Gb of data (Table S1).

Repetitive content and genome size

We estimated the size of the repetitive fraction of each individual's genome based on five satMiner iterations as described in the methods section. The fraction p_i of newly discovered repeats declined as iterations i progressed but stabilized at a positive value (Figure S3). In total, satMiner identified between 2,376 and 5,544 contigs per sample. The fraction of reads q_i that matched repeat clusters increased per iteration and then stabilized (Figure S3). The sum of these two fractions represents an estimate of the total repeat content. This was highly correlated between the two sexes of the same species ($r = 0.96$, $t_4 = 6.56$, $p = 0.0028$) and variable among species, with *biguttulus* showing the lowest repeat content (79%) and *scalaris* the highest (96%, Figure S3). Applying the same procedure to reads from the published *Locusta* genome (Wang *et al.*, 2014) resulted in an estimated repeat content of 71%. Alternative quantifications by a single RepeatExplorer run and based on dnaPipeTE yielded lower, but highly correlated estimates for our set of six species (Table S2).

Genome size quantification was performed using flow cytometry and compared to the three species for which published genome sizes are available (Table S3). Our estimates were similar to previous publications for *scalaris* (13.98 vs. 14.72), lower for *parallelus* (9.73 vs. 12.31) and higher for *sibiricus* (10.43 vs. 8.95). Both these cases might represent population differences, since our measurements were taken from other populations than previous estimates (Table S3). Total repeat content was strongly and positively correlated with genome size across species (gomphocerine species only: $r = 0.87$, $t_4 = 3.62$, $p = 0.022$, including *Locusta*: $r = 0.93$, $t_5 = 5.70$, $p = 0.0023$, Pearson's correlation test, Figure 2).

Characterization of repeat content within species

Averaged across species dnaPipeTE annotated around 24% of the repeatome as DNA transposons, 13% as helitrons, 21% as LINE elements, 12% as LTR retrotransposons, 1.6% as SINE elements, 8.5% as satellite DNA and 19% as low-copy number elements (Figure S4-S5). There was marked variation of the relative proportions of these different repetitive elements among species. Particularly pronounced was the large abundance of satellites in *sibiricus* and *scalaris* and the low abundance of satellites in *parallelus* (Figure S4-S5). Helitrons were found to be quite common in all species, but were most abundant in *scalaris* (Figure S4-S5). Other repeat classes were less variable among species in their relative abundances.

When assembling the repeatome *de novo* using RepeatExplorer, we found a 'tapering' pattern of repeat cluster frequencies in all species and in both sexes (Figure 3). In most species, there was no markedly dominating cluster of repeats. A similar pattern was present in *Locusta* (Figure S6). However, a strikingly different pattern was obtained for *scalaris* as well as for the female *sibiricus* individual, both of which appear to be dominated by a single highly abundant cluster. In these species, the most abundant cluster accounted for around 10–15% of the total number of reads. In all samples of *scalaris*, *sibiricus* and *biguttulus*, as well as in the *variegatus* male, the most abundant cluster was annotated as satellite DNA, while in all other cases the top cluster was either annotated as helitrons or could not be annotated.

Divergences within clusters of transposable elements

We estimated the average divergences within read clusters of transposable element copies using dnaPipeTE (Figure S7-S8). Sequence divergence was highest for SINE elements (6.9%) and DNA transposons (6.3%), intermediate for helitrons (5.9%) and LINE elements (5.4%) and lowest for LTR retrotransposons (4.2%). Variation in sequence divergence across species was low for DNA transposons, LINE elements and LTR retrotransposons, but pronounced for helitrons (lowest in *scalaris*, 4.8%; greater than 5.7% in all other species) and SINE elements (lowest in *scalaris*, 4.5%; greater than 6.5% in all other species).

Variation in repeat content across species

While the sample-by-sample analysis provided an unbiased picture of repeat content distribution within samples, matching clusters across samples was less straightforward. We therefore conducted an additional analysis in which we pooled reads across samples and collectively *de novo* assembled their repeat content. We extracted the first 15 repeat clusters (constituting 12–37% of the genome per sample) and analysed how reads of different samples contributed to these clusters. We found strong positive correlations in repeat content between the two samples from the same species (average Pearson correlation $r = 0.94$ across the first 15 clusters, Figure S9) implying that the two biological replicates within each species were highly similar and that intraspecific differences were low compared to interspecific variation.

To visualise the distribution of repeat clusters both within and among species, we conducted a principle component analysis (PCA) focusing on the 15 most abundant clusters that could be matched across runs. Three main patterns emerged (Figure 4). First, all runs from the same sample clustered tightly together, illustrating that our subsample size was sufficiently large to robustly estimate among-sample variation. Second, samples of females and males from the same species also clustered closely together, except for the two *sibiricus* individuals, which showed a marked intraspecific difference in PC1 values. Third, related species tended to cluster together, in particular the species pair *biguttulus*/*rufus*. To investigate these patterns further, we plotted the frequencies of the most abundant clusters separately for males and females of all species (Figure 5). Variation within *sibiricus* was

found to arise mainly from differences in the abundance of the satellite cluster (cluster 1) although the female also had a higher frequency of cluster 7 (helitrons) and the male had a higher frequency of clusters 6, 9 and 10 (helitrons, LINE1 elements and unnamed, respectively).

Intra-specific differences in *Gomphocerus sibiricus*

The male *sibiricus* sample was unusual in several aspects (cluster size distribution, Figures 3; principle component analysis, Figure 5; sequence divergence within clusters, Table S4). However, three lines of evidence suggest that these patterns were not simply caused by sample mix-up, sequencing artefacts or contamination, since (a) both independent MiSeq and HiSeq runs yielded similar patterns, (b) the samples of the two *sibiricus* individuals clustered together in our phylogenetic reconstruction based on mitochondrial reads (Figure S10-S11), and (c) blast queries against standard databases did not yield any unusual hits. Nevertheless, we placed more confidence in the female *sibiricus* sample because of the better match with independent samples analysed previously (Shah *et al.*, 2016).

For among species comparisons, the characteristic feature of the *rufus*/*biguttulus* pair was the high abundance of helitrons of clusters 2 and 8 and the low abundance of cluster 10. *Scalaris* showed a particularly high abundance of satellites (cluster 1) and helitrons of cluster 7. *Parallelus* and *variegatus* as the two most divergent species in our dataset showed rather different distributions, with *variegatus* being an outlier in the principle component analysis (Figure 4) and *parallelus* in the abundance of clusters 1–4 (Figure 5). *Parallelus* was characterised by a low abundance of satellites (cluster 1) and helitrons of clusters 2 and 7, but a relatively high abundance of helitrons from clusters 4 and 6. *Variegatus* was different in being rather average in representation across clusters. Mapping changes in clusters size across the phylogeny using ancestral state reconstruction provided tentative evidence for increases in satellites (cluster 1), helitrons (cluster 7), simple repeats (cluster 15) and unknown (cluster 3) from the most ancestral species (*parallelus*/*variegatus*) to the most derived species (*sibiricus*/*scalaris*), but also some apparent decreases in cluster sizes, such as for helitrons of cluster 11 (Figure S12). Strongest positive correlations between repeat abundance and genome size were found for cluster 1 (satellite), cluster 7 (helitron) and cluster 15 (simple repeats) (Table S8).

Species differences explored by cluster painting

Reads within clusters (as identified by RepeatExplorer) can be visualized as graphs in which individual reads are represented by nodes and read overlaps by edges. If a given repeat class spread prior to the split of two species, we would expect reads of those species to be distributed randomly across graphs due to sequence divergence prior to and after the species split. By contrast, if a repeat class expanded and diverged after the split of two species, we would expect reads from the same species to cluster together within graphs. We therefore colour-coded reads by sample in the joint graph in an approach that can be described as 'pool-and-paint' cluster painting (Figure 6, Figure S13). We found that clusters 1 (annotated as satellite DNA) showed closer relationships of reads within species as opposed to between species (Figure 6), indicating sequence divergence after species split. Clusters 3 and 7 showed similar tight clustering of reads from *biguttulus* and *rufus* that both covered similar regions of the graph (Figure 6, Figure S13). In contrast, clusters 2, 4-6 and 9-10 showed a much more even distribution of samples across graphs (Figure 6, Figure S13), suggesting that the divergence is older such that diversity is shared among species.

Discussion

We here present a comparative analysis of the repeat content of six species of gomphocerine grasshoppers, including *Stauroderus scalaris*, which has the second largest insect genome described to date (Gregory, 2018). We found a large fraction of retrotransposons, in particular LINEs and LTRs but few SINEs, and a relative high abundance of satellite DNA and helitrons. We also found substantial variation in repeat content among species, while marked intraspecific differences were only found in *Gomphocerus sibiricus*. The distribution across repeat classes was evenly skewed in most of the species, apart from *sibiricus*/*scalaris*, where a single repeat class was dominant, indicative of a recent expansion of satellite DNA in these two species or their common ancestor. The remaining species exhibited a relatively even distribution of repeat classes, suggesting that invasion by repeats is either ancient or that multiple repeat types spread simultaneously in the more recent past. The latter conclusion is supported by the relatively young and unimodal distribution of divergence times within clusters.

Repeat content varied between 79 and 87% across most of the species, the only exception being *scalaris*, which had an estimated repeat content of 96%. Overall, there was a strong positive correlation between repeat content and genome size as described elsewhere (Charlesworth *et al.*, 1994; Petersen *et al.*, 2019; Talla *et al.*, 2017; Wu & Lu, 2019). The repeat content in *Locusta* (genome size 6.44 pg) was estimated at 71% using our method, which linearly prolongs the positive correlation between genome size and repeat content. Repetitive elements are thus likely drivers for genome size expansion, possibly due to positive feedbacks that allow these elements to spread more easily in large genomes (Hollister & Gaut, 2009). Our asymptotic estimate of repeat content in *Locusta* was slightly higher than that of Wang *et al.* (2014), possibly reflecting the difficulty of assembling and estimating repeat content through genome assembly (Wang *et al.*, 2014).

One of our most striking results was the expansion of satellite DNA in *sibiricus*/*scalaris*. We suggest that causality might be reversed in this case, in the sense that satellite DNA may not be the cause of genome size expansion, but rather a consequence. Previous studies suggest that satellite DNA may contribute substantially to genome size in grasshoppers with large genomes (Ruiz-Ruano *et al.*, 2016; Shah *et al.*, 2016). Satellite DNA is known to be particularly abundant in in the centromeric and telomeric parts of the genome and leads to densely packed heterochromatin structures (Plohl *et al.*, 2008). Centromeric heterochromatin has a function in the pairing of sister chromatids and is therefore important for proper cell division (Hartl, 2000; Plohl *et al.*, 2008). It is conceivable that a stabilizing function of satellite DNA might be required when chromosomes become greatly expanded as in the case of grasshoppers. Satellite DNA often evolves in a concerted fashion (Garrido-Ramos, 2017; Palomeque & Lorite, 2008; Plohl & Meštrović, 2012), as indicated in our data by the clustering of reads within species, but different variants of satellite motifs seem to be recruited from a conserved pool of ancestral satellites. Satellite DNA occurs both unclustered and spatially clustered in the genome and it has been suggested that local clusters may have evolved secondarily (Palacios-Gimenez *et al.*, 2017; Ruiz-Ruano *et al.*, 2016). If satellite DNA contributes to chromosome integrity, such expansions might be adaptive in species with large genomes.

Our results also suggest that helitrons have accumulated in gomphocerine grasshoppers. Helitrons spread via rolling circle replication (Thomas & Pritham, 2015). They can occur in

large numbers (such as in some plants, (Xiong *et al.*, 2014)) but tend to be rarer than retrotransposons in most animals (Kapitonov & Jurka, 2007). Although we also detected many retrotransposons, the relatively high abundance of helitrons in grasshoppers is noteworthy. As with satellite DNA, it is possible that the abundance of helitrons is not the primary cause of genome size expansion, but that they have proliferated in already large genomes. However, relatively high sequence divergence suggests a relatively old age for the spread of helitrons. *Scalaris* represents an exception to the otherwise largely similar representation across species in that helitrons are particularly common in this large-genome species. There are multiple avenues for such positive feedbacks, including more target insertion sites and weaker negative selection per insertion (Hollister & Gaut, 2009). Helitrons are biologically significant because they often include fractions of non-helitron DNA, sometimes entire genes, and thus offer a vehicle for the genomic translocation of functional elements (Thomas & Pritham, 2015). Furthermore, helitrons and a number of other transposable elements have been shown to be involved in horizontal gene transfers across insects (Peccoud *et al.*, 2017; Wu & Lu, 2019).

In order to visualize interspecific patterns, we mapped species-specific reads to clusters. We used an approach that we describe as pool-and-paint cluster painting to visualize if reads from different samples occupy different parts of the graphs of pooled reads. As we describe above, we pooled reads in order to avoid biases that could arise if we had clustered different libraries independently. Our approach allows shared clusters to appear in the joint analysis even if cluster sizes are small in individual samples. Cluster painting allows explorative assessment, based on the idea that within clusters, reads originating from a recent expansion within a species should cluster more closely together. While this represents an explorative analysis that does not in itself yield a quantitative measure of variation within and among samples, it has the potential to serve as a visualization technique and explorative tool for other applications, particularly when comparing different populations or species. The method relies on sequence differences among lineages and is thus likely to work best for data from rather divergent forms.

Our cluster painting approach showed that reads within cluster graphs were structured by phylogenetic relatedness in at least some cases (Figure 6, Figure S13). This suggests that repetitive elements often proliferated after lineage splits. However, not all clusters showed

such a pattern (e.g. clusters 2, 4–6 and 9-10), suggesting that some elements may have expanded during the earlier phylogenetic history of the Gomphocerinae. The relatively similar sequence divergence within clusters (Figure S7–S8) is also suggestive of older expansions, except for LTR retrotransposons, which appear to be younger (Table S4).

Gomphocerus sibiricus was the only species for which the distribution of repeats differed markedly between the two samples. In principle, this difference may be driven by the sex chromosomes. *Sibiricus* has three large and five medium-sized pairs of autosomes and the X chromosome is of similar size to the smaller autosomes (Gosalvez & López-Fernandez, 1981). It also has an X0 sex determination system, in which females have two and males one copy of the X chromosome. However, as the repeat content of the two sexes did not differ substantially from one another in any of the other species, we consider a sex chromosome explanation unlikely. Alternatively, inter-individual differences within species may result from the presence or absence of supernumerary chromosomes (B chromosomes) or supernumerary segments of normal chromosomes, which are facultatively present in some individuals (Gosalvez & López-Fernandez, 1981). However, the male *sibiricus* sample was unusual in several aspects and also differed markedly from data generated for different individuals of the same species in a recent study (Shah *et al.*, 2016). Consequently, it is possible that this particular sample may be untypical, possibly due to genuine differences in genome structure, or alternatively as a result of unknown biases that could have arisen during the sequencing or assembly procedure. However, the congruence of the two independent library preparations and sequencing runs as well as the results of our mitochondrial phylogenetic reconstruction suggest that these differences probably have a biological rather than technical origin.

Overall, our analysis of repeat content in the large genomes of gomphocerine grasshoppers reveals a strong link between genome size and repeat content, and in particular high abundances of various helitrons and satellite DNA. We suggest that the expansion of satellite DNA might be secondary and could potentially have been favoured by selection as a means of stabilising these greatly expanded genomes. Whether or not helitrons played a primary or secondary role in grasshopper genome size expansions remains an open question, but it seems reasonable to speculate that increases in genome size likely followed

a multi-step process, in which different repetitive elements proliferated during the earlier and later phases of genome size expansion.

Methods

Species and sample collection

We sampled hind legs from one male and one female each of six species from the subfamily Gomphocerinae of acridid grasshoppers (total *n* = 12 individuals): meadow grasshoppers *Pseudochorthippus parallelus* (Bielefeld, Germany), alpine thick-necked grasshopper *Aeropedellus variegatus* (Engadin, Switzerland), rufous grasshopper *Gomphocerippus rufus* (Engadin, Switzerland), bow-winged grasshopper *Chorthippus biguttulus* (Bielefeld, Germany), club-legged grasshopper *Gomphocerus sibiricus* (Engadin, Switzerland) and large mountain grasshopper *Stauroderus scalaris* (Engadin, Switzerland). Based on previous mitochondrial analyses (Dumas *et al.*, 2010; Vedenina & Mugue, 2011) as well as our own results (Figure 1), *sibiricus-scalaris* and *biguttulus-rufus* appear to be sibling taxa, while *parallelus* and *variegatus* are more distantly related. Hind legs were stored in 70% ethanol at -20°C prior to DNA extraction from postfemur muscle tissue using a standard chloroform-isoamyl alcohol extraction protocol (Sambrook *et al.*, 1989).

Genome size determination by flow cytometry

We quantified genome sizes by flow cytometry following a standard protocol (Hare & Johnston, 2011). Nuclei were extracted from heads of three male grasshoppers per species. Preliminary analyses have shown that freezing after nuclei isolation leads to blurred peaks in the flow cytometer. Therefore, all samples were processed immediately before measurement. Half a brain, split longitudinally, was used per extraction. First, 1 ml of cold Galbraith buffer was added to each sample. Samples were then ground with 15 strokes of a pestle in a Dounce grinder. Both the grinder and pestle were washed with Milli-Q water between the processing of each sample. Homogenates were transferred to Eppendorf tubes and left to incubate for 15 minutes. Ground samples were filtered through a 20 µm nylon mesh filter to remove cell debris and the filtrate was recovered into a 5 ml falcon tube on ice. 20 µl (5% of the total volume) of the standard *Acheta domesticus* extract was added to

each sample. Each extract was further diluted with 100 µl of 0.5 mg/ml propidium iodide to obtain a final concentration of 50 µg/ml. Samples were left to stain for one hour on ice in the dark before being filtered again using a 20 µm nylon mesh filter and then analysed on a BD FACS Canto II flow cytometer. Analyses continued at a medium flow rate until 10,000 gated events were recorded.

Flow cytometry data were processed using the BD FACSDiva software. Besides the pronounced peak of the cricket size standard, we usually observed a smaller peak at approximately twice the signal intensity that was putatively caused by mitotically dividing cells. A second peak at twice the signal intensity of the target sample was also sometimes visible, but the peak was small and usually blurred, so that it could not be analysed. However, these results demonstrate overall linearity of the signal across the observed range. We converted signal intensities to genome sizes by taking the least squares fit of published genomes sizes (averages available for four species, Table S3) on signal intensity (adjusted R^2 = 0.82, Figure S14).

High throughput sequencing and short read pre-processing

We generated separate sequencing libraries for all 12 individuals using an Illumina Nextera DNA library preparation kit and size-selected fragments ranging from 300 to 700 bp. These libraries were then 2x300 bp paired-end sequenced on the Illumina MiSeq sequencing platform, which resulted in 4.5 Gb of sequence and an average depth of coverage across the entire genome of around 0.0034x. To further increase the quantity of data, we sequenced the same samples with 150 bp single-end reads on two Illumina HiSeq 2500 lanes to yield 31.1 Gb of sequence, corresponding to an average depth of coverage of around 0.23x. The resulting raw reads were pre-processed and filtered using trimmomatic (version 0.36, (Bolger *et al.*, 2014)) and FASTX toolkit (version 0.06, (Gordon & Hannon, 2010)) to remove sequencing adapters, sequencing artefacts and low quality reads (<20 phred). Trimmomatic was set to remove sequencing adapters, leading and low quality bases (below quality 3), bases which fall below quality 15 in a 4 bp wide window and reads with final lengths below 120 bp.

Phylogenetic analysis

We used MitoFinder (version 1.2,(Allio *et al.*, 2020)), a pipeline to extract and assemble mitochondrial genome from sequencing data, to harvest as many mitochondrial sequences as possible from all samples. Although nuclear sequences would be preferable for phylogenetic reconstruction, our low coverage sequencing does not yield sufficient coverage of well-represented nuclear genes. Nevertheless, mitochondria are present in higher copy numbers than nuclear mitochondrial copies (which frequently cause problems for phylogenetic analysis in orthopterans, (Hawlitschek *et al.*, 2017; Song *et al.*, 2014)) and are therefore ideally suited for phylogenetic analysis. We used MAFFT (version 7.313, (Katoh & Standley, 2013)), with the L-INS-i option to create a multiple sequence alignment of mitochondrial genes. We reconstructed phylogenies on a gene-by-gene basis for 15 mitochondrial genes (Figure S11). Since many genes had missing sequences for some samples, we selected the COI, COII and COIII genes, which had the least missing data, for a final analysis in which multiple sequence alignments were concatenated (Figure S10). *Pacris xizangensis* (Li *et al.*, 2020) was added as an outgroup for rooting. The phylogenetic analysis was performed using PartitionFinder (version 2.1.1, (Lanfear *et al.*, 2016)) in order to select best-fitting partitioning schemes and models of molecular evolution, followed by a maximum-likelihood based phylogeny estimating using RAxML (version 8.2.12, (Stamatakis, 2014)), with a GTR substitution model and GAMMA rate heterogeneity across sites. *De novo* repeat identification

De novo repeat identification

We used RepeatExplorer (version 0.9.7.8) for *de novo* repeat identification (Novák *et al.*, 2013). Clustering was based on read similarity across multiple copies of repeat elements and in the ideal case, clusters represent all reads from a family of repeats. RepeatExplorer relies on RepeatMasker (version 4.06, (Smit, 2015)), RepBase (version 20160829, (Bao *et al.*, 2015)) and Dfam (version 2.0, https://dfam.org/help/tools) for identification of repeat families. Initially we did this separately for each sample based on HiSeq reads. As RepeatExplorer can handle only a limited number of reads, we randomly selected 10% of the reads from each sample. This process was repeated five times but the replicate runs

yielded virtually identical results, so we present only data from a single RepeatExplorer run per sample (Figure 3).

We conducted an independent analysis to confirm our results from RepeatExplorer using dnaPipeTE (version 1.3, (Goubert *et al.*, 2015a)), an alternative pipeline for the *de novo* assembly, annotation and quantification of transposable elements. We ran dnaPipeTE with default settings and five Trinity iterations. dnaPipeTE is a fully automated pipeline to assemble and quantify repeats, which assembles repeats from short-read data using the Trinity *de novo* transcriptome assembler in an iterative fashion. This is followed by annotation of the assembled contigs using RepeatMasker and the RepBase database. Finally, BLASTN is used to estimate the relative abundance of transposable elements, to shed light on the transposable element divergence landscape, and to further annotate the assembled unannotated contigs.

Iterative repeat identification and filtering

We used a custom version of satMiner (Ruiz-Ruano *et al.*, 2016) to filter the sequence data for reads associated with repetitive elements and to estimate the total repeat content per sample. The 12 libraries and the MiSeq and HiSeq reads were processed separately at this stage, resulting in 24 satMiner runs. satMiner uses RepeatExplorer to analyse a small subset of each library (set to 300,000 reads) in order to identify repeat clusters *de novo*. The fraction of reads assigned to repeat clusters was then used to query the remainder of the sequences. Sequences of high similarity were assigned to newly identified clusters and removed from the pool of sequences before progressing with the next iteration of satMiner by parsing a new subset of 300,000 reads from the remaining pool of reads to RepeatExplorer.

We ran satMiner for five iterations, which involved six *de novo* assembly steps and five mapping and filtering steps. As satMiner does not retain reads which are assigned to clusters, we modified the code so that this information was retained. Our modified version of satMiner is available via https://github.com/abshah/satminer. To facilitate downstream analyses, the MiSeq read pairs were merged using PEAR (version 0.9.10, (Zhang *et al.*, 2014)). We then used custom Linux shell scripts to collate MiSeq and HiSeq reads revealing

homology to repeat clusters identified by satMiner into a single readsets, which we refer to as 'repeat-enriched readsets'.

Again, we used the dnaPipeTE pipeline as an independent method to analyse repeat-enriched readsets. We ran dnaPipeTE with default settings with the number of Trinity iterations set to 5 on all repeat-enriched readsets. Results of repeat-enriched readsets were similar to the dnaPipeTE analysis of full readsets before enrichments (see above) and we therefore present only the former.

Repeat content estimation

The five successive satMiner iterations were used to estimate the total repeat content of each sample. During each iteration *i*, we quantified the percentage of the reads that was *de novo* assigned to clusters, p_i. We then searched for the set of reads q_i that showed sequence similarity to reads in p_i. As reads that are assigned to clusters (p_i) or that show sequence similarity to reads within clusters (q_i) were sequentially removed, we expected this fraction to decline progressively with each iteration. However, we found that p_i remained approximately constant across iterations, while querying the remaining pool of reads gave rapidly diminishing yields of repetitive sequences q_i (Figure S3). This suggests that the query step was not fully efficient and that each iteration re-discovered the same repeat clusters rather than finding new ones. In fact, the sum of the fraction filtered out of the total pool and the fraction assigned *de novo* to clusters quickly stabilized after two iterations (Figure S3). We therefore used the sum $\Sigma(p+q)$ calculated after the last satMiner iteration to provide the best estimate of total repeat content.

Joint repeat clustering and comparison across species

Comparing clusters across species can sometimes be difficult due to issues with merging clusters across independent runs in different readsets. Consequently, we analysed readsets that contained reads from different individuals and species in equal proportions as described below. We processed the repeat-enriched readsets using RepeatExplorer (version 0.9.7.8, (Novák *et al.*, 2013)). In order to ensure equal representation of repetitive elements from all biological samples, we sub-sampled each of the twelve enriched readsets 20 times

without replacement, each time drawing 25,000 MiSeq reads and 75,000 HiSeq reads at random to produce a total sub-sample of 100,000 reads per readset. This generated 20 datasets, each comprising 1,200,000 sub-sampled reads pooled over all 12 individuals that were analysed by RepeatExplorer to generate *de novo* assembled repeat clusters.

We then used reciprocal BLAST to match contigs from clusters identified by RepeatExplorer pairwise across independent runs. We aimed to pool the 15 most abundant repeat classes that we assumed to be represented in all runs. As rank order may change across runs, we used the first 50 clusters produced by each run to determine pairwise matches (of which the first 30 are shown in Table S5). Within the pool of 50x50 reciprocal BLAST matches across 50 clusters from each of two runs, there was a single best match for the most abundant 15 clusters in all cases (Table S6). Reads from clusters identified as best matches were pooled and the 15 clusters with the most reads across pooled samples were further processed.

We used principle component analysis (PCA) to compare the overall pattern of repeat clusters across individuals. This was based on the 15 most abundant clusters keeping the 20 replicated sampling draws as independent cases as they contained no overlapping reads. The PCA was therefore performed on 15 items (clusters) and 240 cases (20 replicated subsamples each of 12 individuals). We performed the PCA with variance-standardized items, thus giving all clusters equal weight in the analysis. The first three axis showed eigenvalues above unity and thus explained more variance than any of the original clusters alone. Analyzing only the first ten clusters yielded qualitatively similar results (with two eigenvalues above unity).

Furthermore, we identified reads from different biological samples by visualizing aggregations of reads from different species in different regions of the cluster graphs. Cluster graphs were built on the repeat-enriched pool across all samples and we thus refer to this approach as ‘pool-and-paint’ cluster painting.

Cluster annotation

Cluster contigs were annotated by RepeatMasker using the Metazoan database of repeats from RepBase (version 20160829, (Bao *et al.*, 2015)). dnaPipeTE uses RepeatMasker and RepBase database for annotation and we used BLASTN to further annotate the assembled

unannotated contigs (Table S7). Annotating *de novo* assembled clusters is challenging and not all annotations are likely to be correct. Nevertheless, most of our analyses relied on relative cluster sizes and the distribution of reads from clusters across samples, and so were not dependent on accurate annotations.

Ancestral state reconstruction

We used ancestral state reconstruction to estimate changes in repeat abundances separately for the major repeat clusters in our set of species. Topology and branch lengths were based on our mitochondrial phylogenetic tree. Repeat abundance was estimated from our RepeatExplorer analysis by multiplying the proportion of reads assigned to each cluster with the estimated genome size of each species. This resulted in an estimate of total sequence content per cluster for each sample. Estimates for males and females were highly correlated and were therefore averaged in the analysis. We then implemented ancestral state reconstruction using REML fits based on a Brownian motion model (as implemented in the ace function of R package ape, version 5.3, (Paradis & Schliep, 2019)) to estimate ancestral states for each node. These were subsequently converted to changes per branch in Mb of sequence per haploid genome.

Comparative analysis of the migratory locust

For some of our analyses, we also incorporated published sequence data from the migratory locust *L. migratoria*, the only acridid species (from the subfamily Oedipodinae) for which a draft genome has been published (Wang *et al.*, 2014). Raw paired-end Illumina HiSeq 2000 sequences (73.6 Gb) were downloaded from the short read archive (accession number SRR764584 and SRR764591). We merged read pairs using PEAR (version v0.9.10, (Zhang *et al.*, 2014)) to create a readset with long single-end reads for comparability with our analysis of gomphocerine species described above. Merged reads below 60 bp were removed. We did not combine reads from *L. migratoria* with reads from the six gomphocerine species in our pooled RepeatExplorer analysis because the species is too distantly related and would distort the pattern of interspecific variation.

Data access

The short read data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under the accession number PRJNA559340. Additional files have been deposited in the iDiv data repository (http://idata.idiv.de/ddm/Data/ShowData/1838).

Tables and Figures

Figure legends

Figure 1: Phylogenetic relationships among six species of gomphocerine grasshoppers (tree rooted using *Pacris xizangensis* as an outgroup) with *Locusta migratoria*, a species used for comparison in some analyses, added with unestimated branch length (the divergence time from gomphocerine grasshoppers is about 61 Mya, (Song *et al.*, 2015)). This phylogeny was based on mitochondrial markers (using COI, COII and COIII genes). Numbers show branch lengths and pie charts at nodes show bootstrap support. The topology is congruent with COI mitochondrial sequence-based analyses published by Vedenina and Mugue (2011) and Dumas *et al.* (2010).

Figure 2: Relationship between repeat content as estimated by *de novo* clustering (see Figure S3) and genome size as estimated by flow cytometry (see Figure S2) for six species of gomphocerine grasshoppers and *Locusta migratoria*. par = *Pseudochorthippus parallelus*, var = *Aeropedellus variegatus*, ruf = *Gomphocerippus rufus*, big = *Chorthippus biguttulus*, sib = *Gomphocerus sibiricus*, sca = *Stauroderus scalaris*, mig = *Locusta migratoria*.

Figure 3: Distribution of *de novo* assembled repeat content over repeat clusters. The upper half of the plot shows results for the female sample while the lower half shows the male sample. Each histogram is based on a single clustering run, with other runs being qualitatively similar. Dashed vertical lines show the estimated repeat content for males and females as estimated by RepeatExplorer based on this single run.

Figure 4: Principle component analysis of repeat content (based on the 15 most abundant clusters) across six species of gomphocerine grasshoppers using variable scaling and rotation of axes. The first three principle components explain 48%, 25% and 15% of the

variation respectively. Each point represents the results of a single run, with species distinguished by colour, females shown as circles and males as triangles. par = *Pseudochorthippus parallelus*, var = *Aeropedellus variegatus*, ruf = *Gomphocerippus rufus*, big = *Chorthippus biguttulus*, sib = *Gomphocerus sibiricus*, sca = *Stauroderus scalaris*.

Figure 5: Abundance of the ten most abundant repeat clusters across six species of gomphocerine grasshoppers. Species are arranged horizontally according to their phylogenetic relatedness, as shown in Figure 1. Females are shown in black and males are shown in grey. Each dot represents one of twenty independent clustering runs based on non-overlapping subsets of the data. par = *Pseudochorthippus parallelus*, var = *Aeropedellus variegatus*, ruf = *Gomphocerippus rufus*, big = *Chorthippus biguttulus*, sib = *Gomphocerus sibiricus*, sca = *Stauroderus scalaris*.

Figure 6: Cluster-paining approach to species-specific differences within cluster. The plot shows the four largest cluster with dots representing reads and read overlap by edges. The six different species are shown by different colours. Tight clustering of reads from the same species (as for clusters 1 and 3) indicate divergence within a species, while dispersion of colours across the graph (as for clusters 2 and cluster 4) indicates that either cluster expansion predates divergence or expansion has continued from a range of diversified repeat copies.

Figure 1

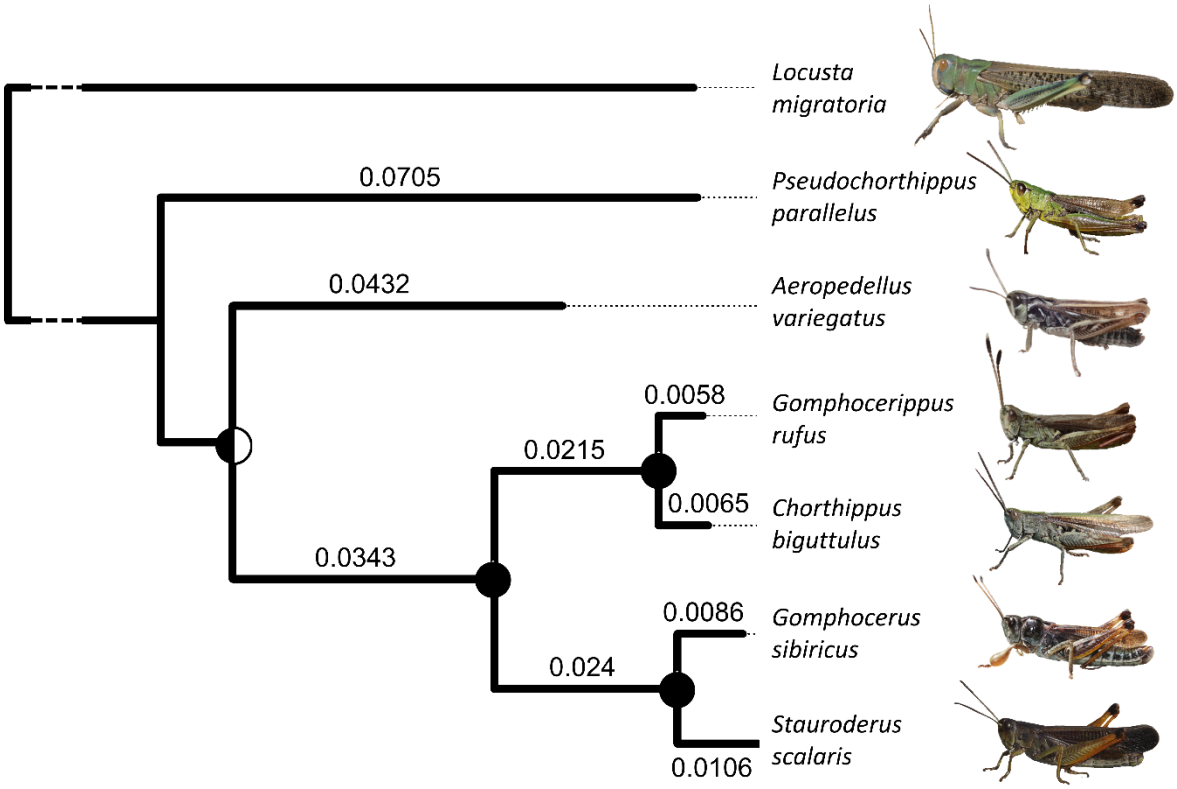

Locusta migratoria
0.0705
Pseudochorthippus parallelus
0.0432
Aeropedellus variegatus
0.0058
Gomphocerippus rufus
0.0215
0.0065
Chorthippus biguttulus
0.0343
0.0086
Gomphocerus sibiricus
0.024
Stauroderus scalaris
0.0106

Figure 2

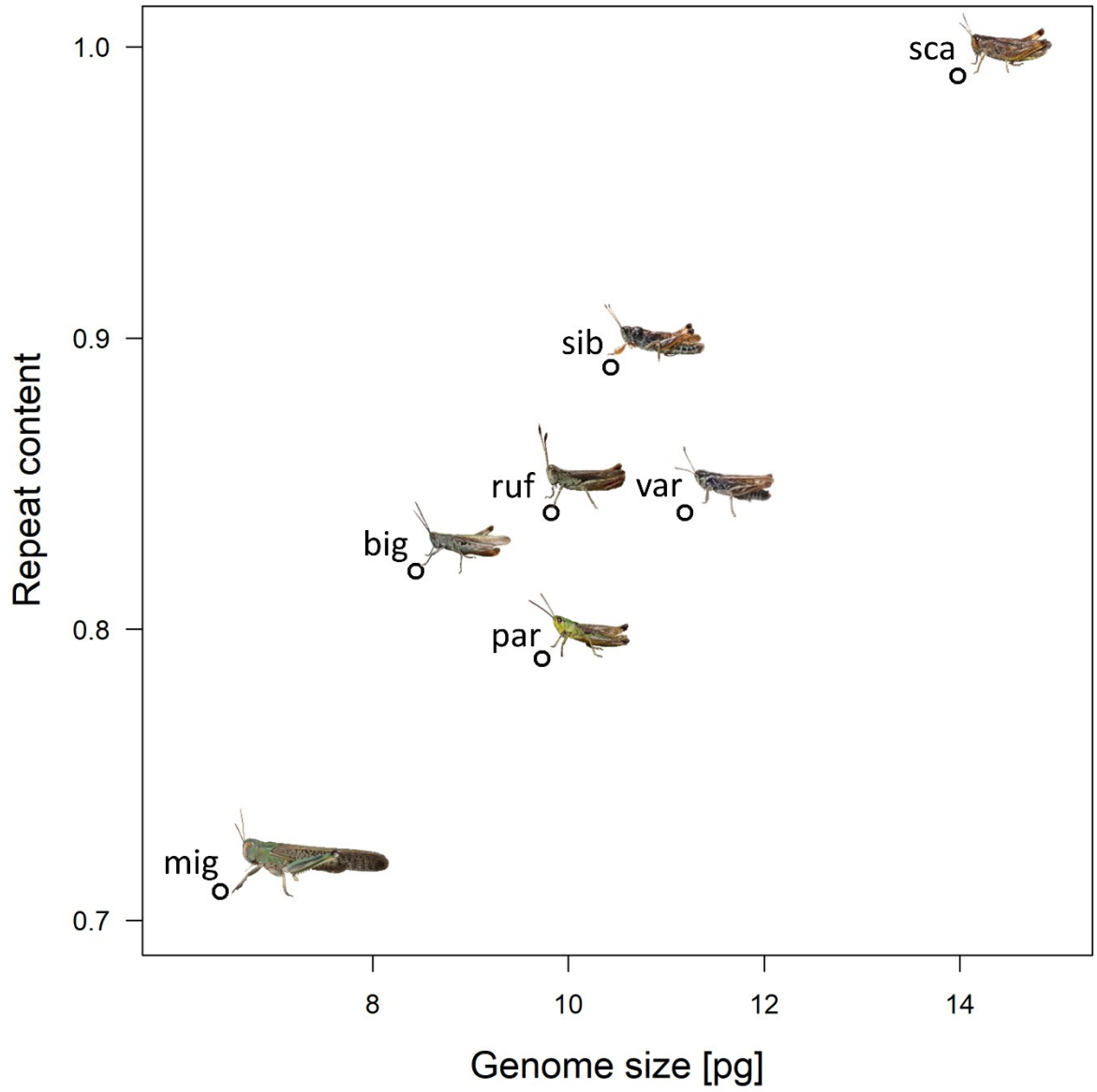

Repeat content
1.0
0.9
0.8
0.7
8
10
12
14
Genome size [pg]
sca
sib
ruf
var
big
par
mig

Figure 3

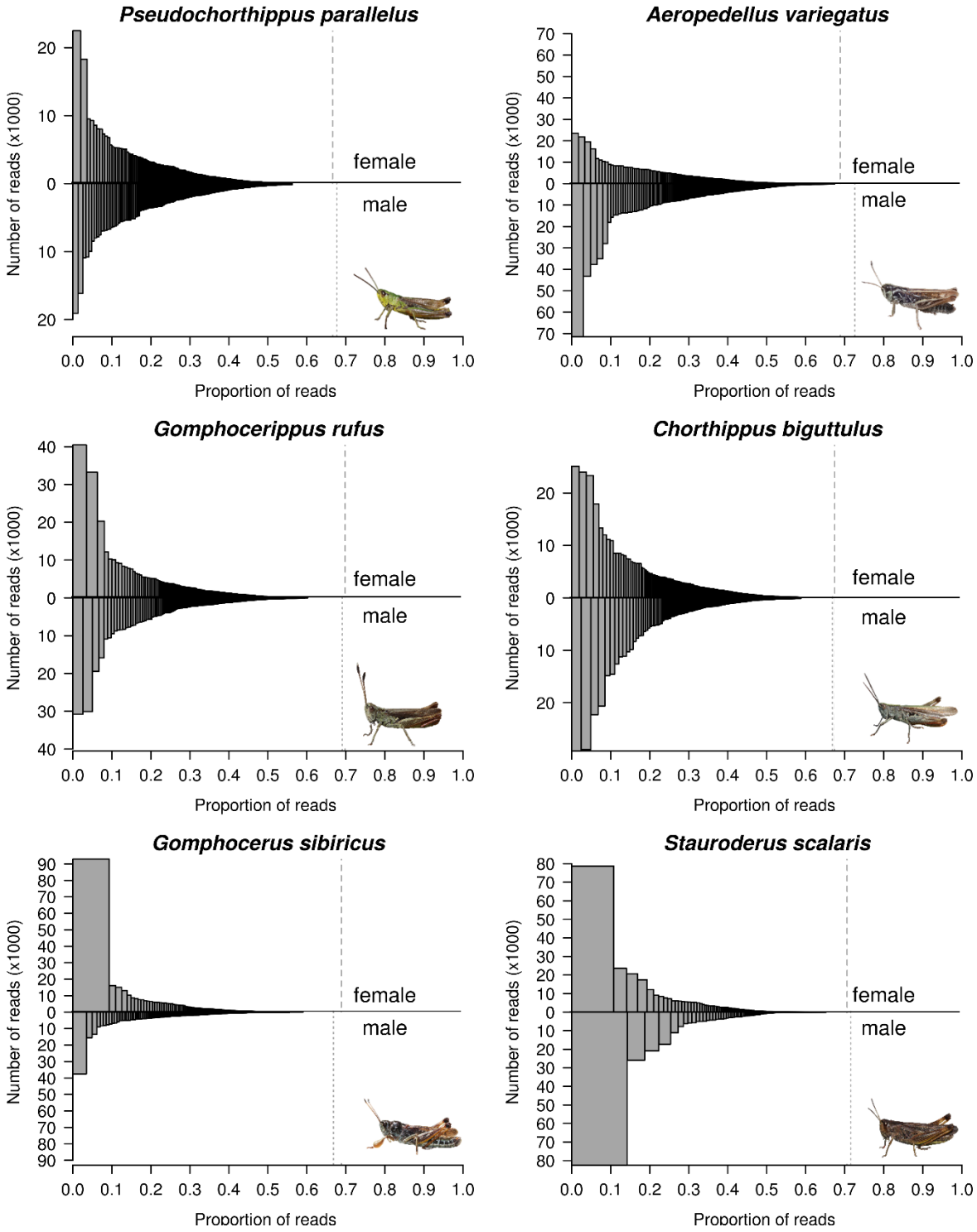
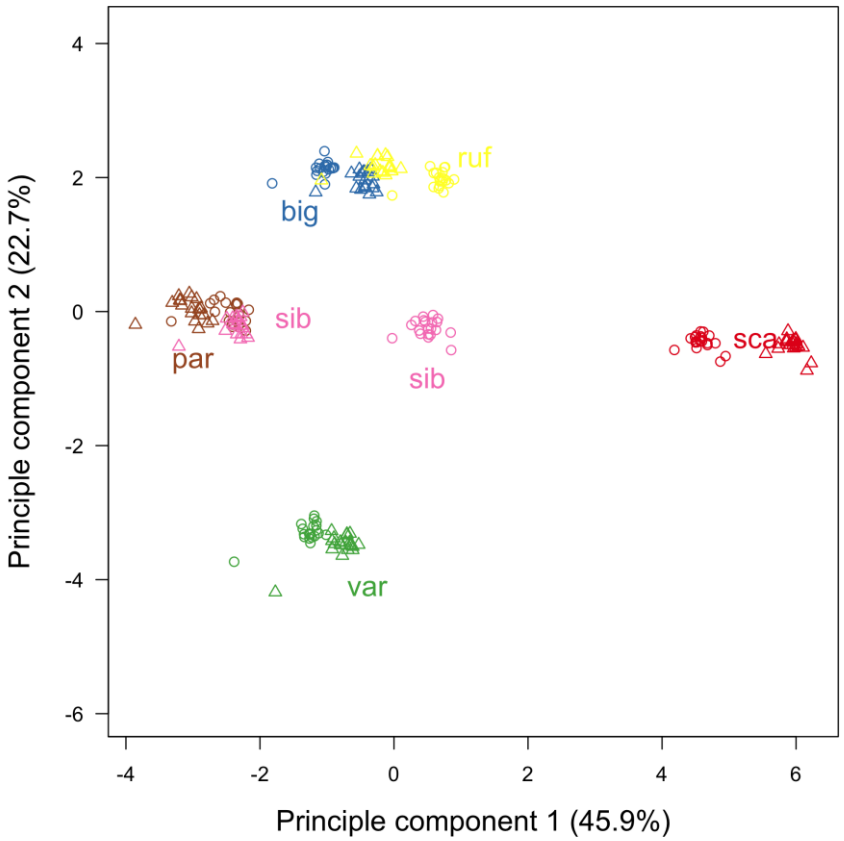

Pseudochorthippus parallelus
Aeropedellus variegatus
Gomphocerippus rufus
Chorthippus biguttulus
Gomphocerus sibiricus
Stauroderus scalaris
Number of reads (x1000)
Proportion of reads
female
male

Figure 4


Principle component 2 (22.7%)
Principle component 1 (45.9%)
big
ruf
sib
par
sib
sca
var
4
2
0
-2
-4
-6
-4
-2
0
2
4
6

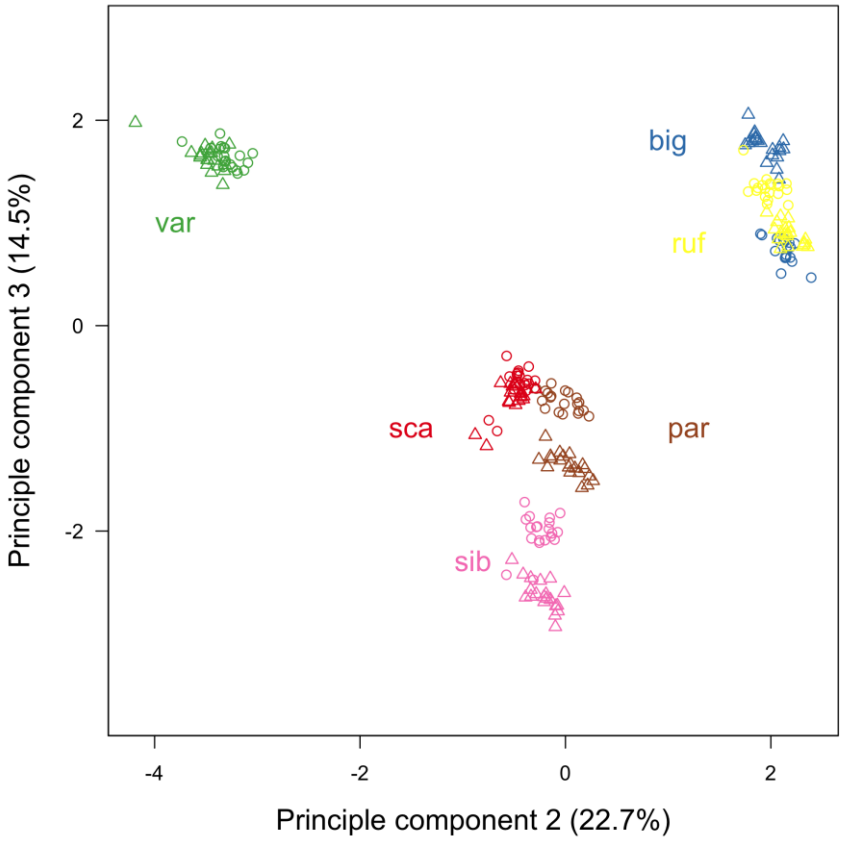

Principle component 3 (14.5%)
Principle component 2 (22.7%)
var
big
ruf
sca
par
sib
2
0
-2
-4
-2
0
2

Figure 5

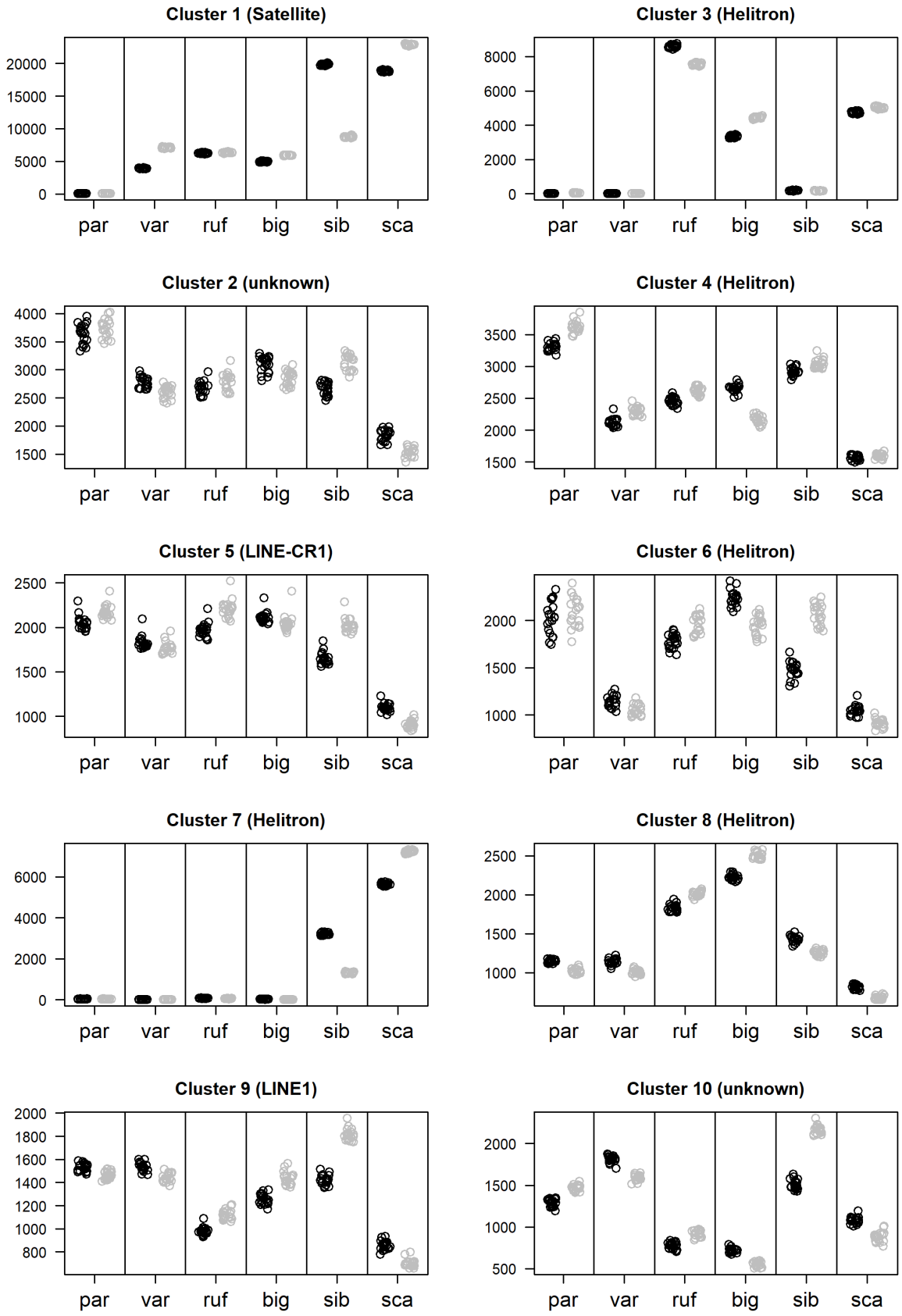
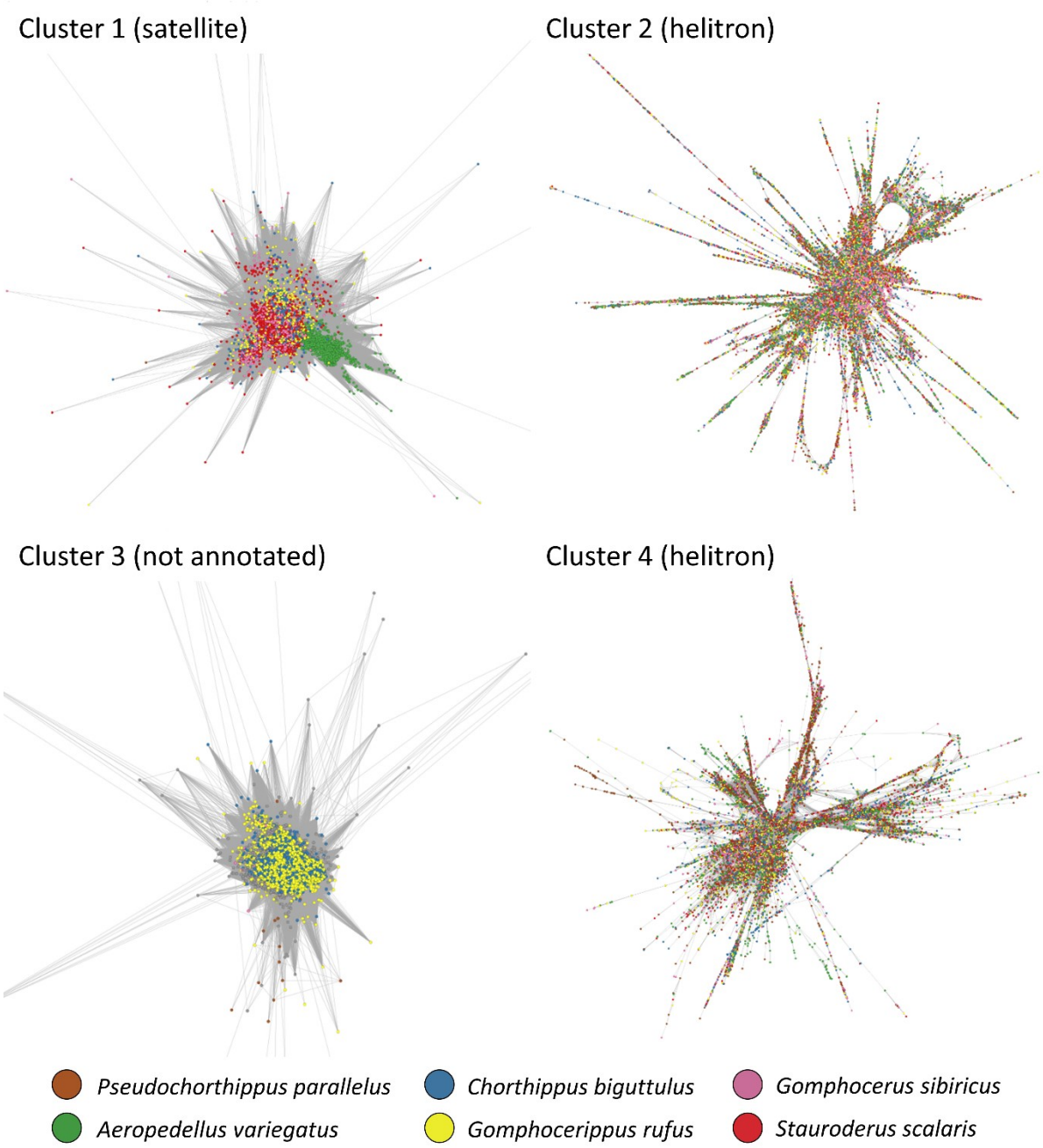

Cluster 1 (Satellite)
Cluster 3 (Helitron)
Cluster 2 (unknown)
Cluster 4 (Helitron)
Cluster 5 (LINE-CR1)
Cluster 6 (Helitron)
Cluster 7 (Helitron)
Cluster 8 (Helitron)
Cluster 9 (LINE1)
Cluster 10 (unknown)
par
var
ruf
big
sib
sca

Figure 6


Cluster 1 (satellite)
Cluster 2 (helitron)
Cluster 3 (not annotated)
Cluster 4 (helitron)
Pseudochorthippus parallelus
Chorthippus biguttulus
Gomphocerus sibiricus
Aeropedellus variegatus
Gomphocerippus rufus
Stauroderus scalaris

Supplementary Material for Manuscript III

Figure S1: Flow chart of the data processing procedure. Each species is symbolized by a different shade of grey and the two sexes are symbolized by filled and hatched texture. Horizontal lines symbolize subsets of the data.

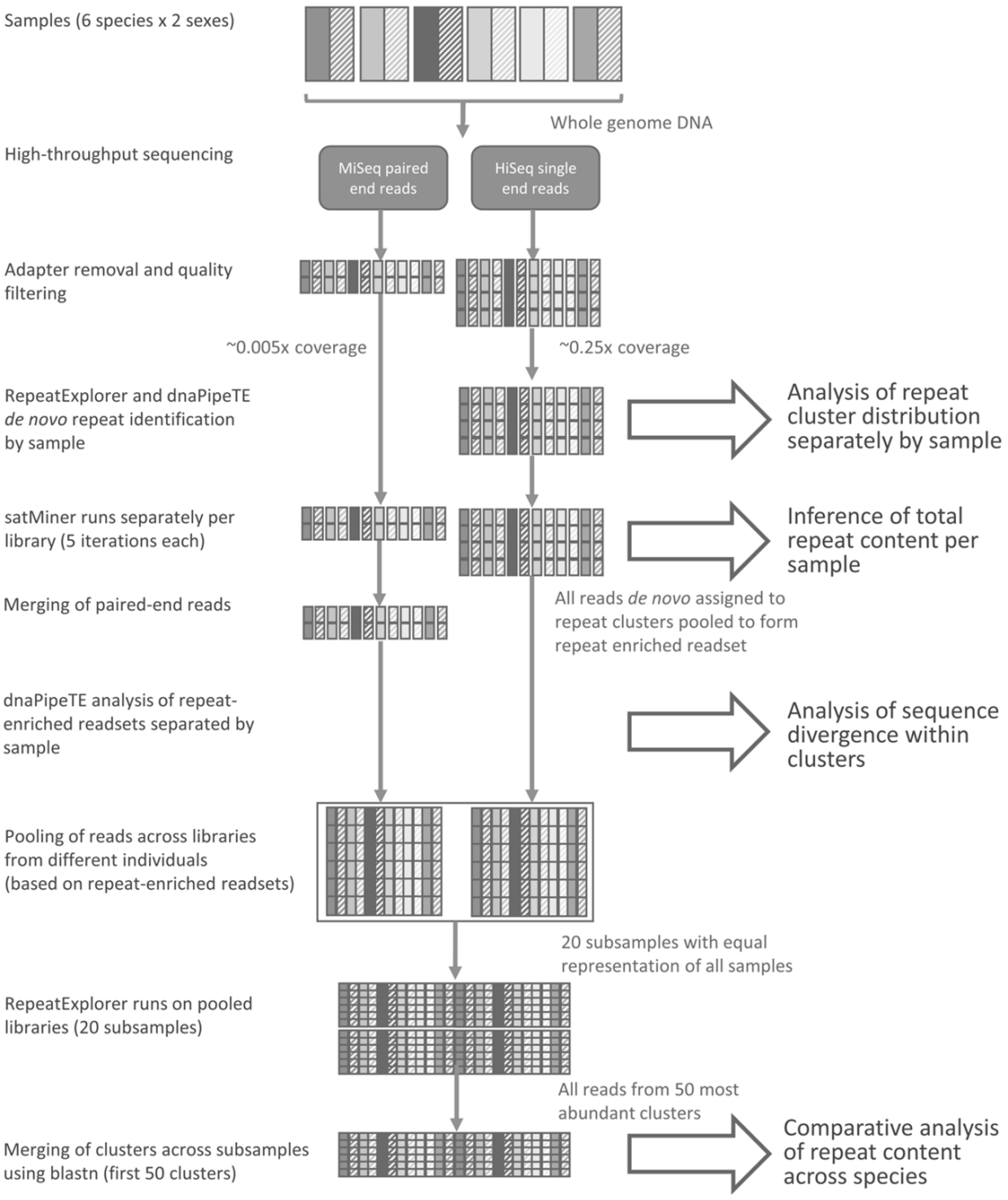


Figure S2: Genome size as determined by flow cytometry based on three males per species using a single *Acheta domesticus* male (diamonds) as a size standard (horizontal line shows the mean value across all *Acheta domesticus* measurements). Flow cytometry signals were converted to genome sizes by regression (Figure S23). rCV varied between 3.8 and 7.4 per sample with a mean of 3.9-7.0 per species (lowest in *Chorthippus biguttulus*, highest in *Gomphocerus sibiricus*). Numbers above data points show average values (± SE) per species.

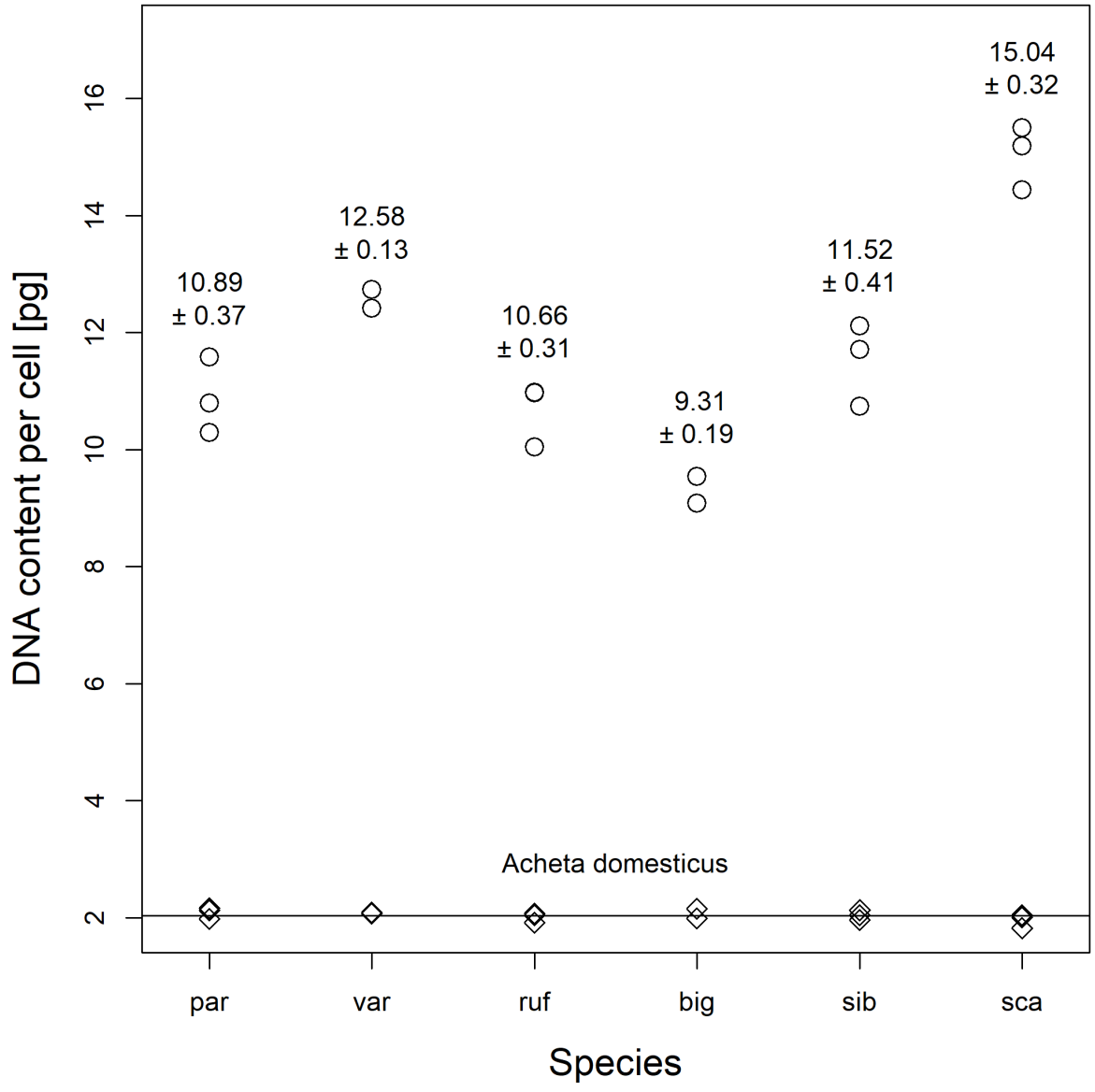


Figure S3: Repeat content estimation based on multiple satMiner iterations. Dashed lines show the proportion of reads (p_i in main text) *de novo* assigned to repeat clusters in each iteration *i*. Dotted lines show the proportion of reads identified as repeats (based on *de novo* clustering of a subset in combination with querying the main pool, q_i in main text). Solid lines show the sum $p_i + q_i$ of the proportion of reads already identified as repeats and the fraction identified as repeats in a subset of the remaining reads. Numbers in the upper left corner show the cumulative estimate of repeat content in the final iteration. Numbers in the lower right corner show the total number of contigs assembled by satMiner. Results for females are shown in black and results for males in grey.

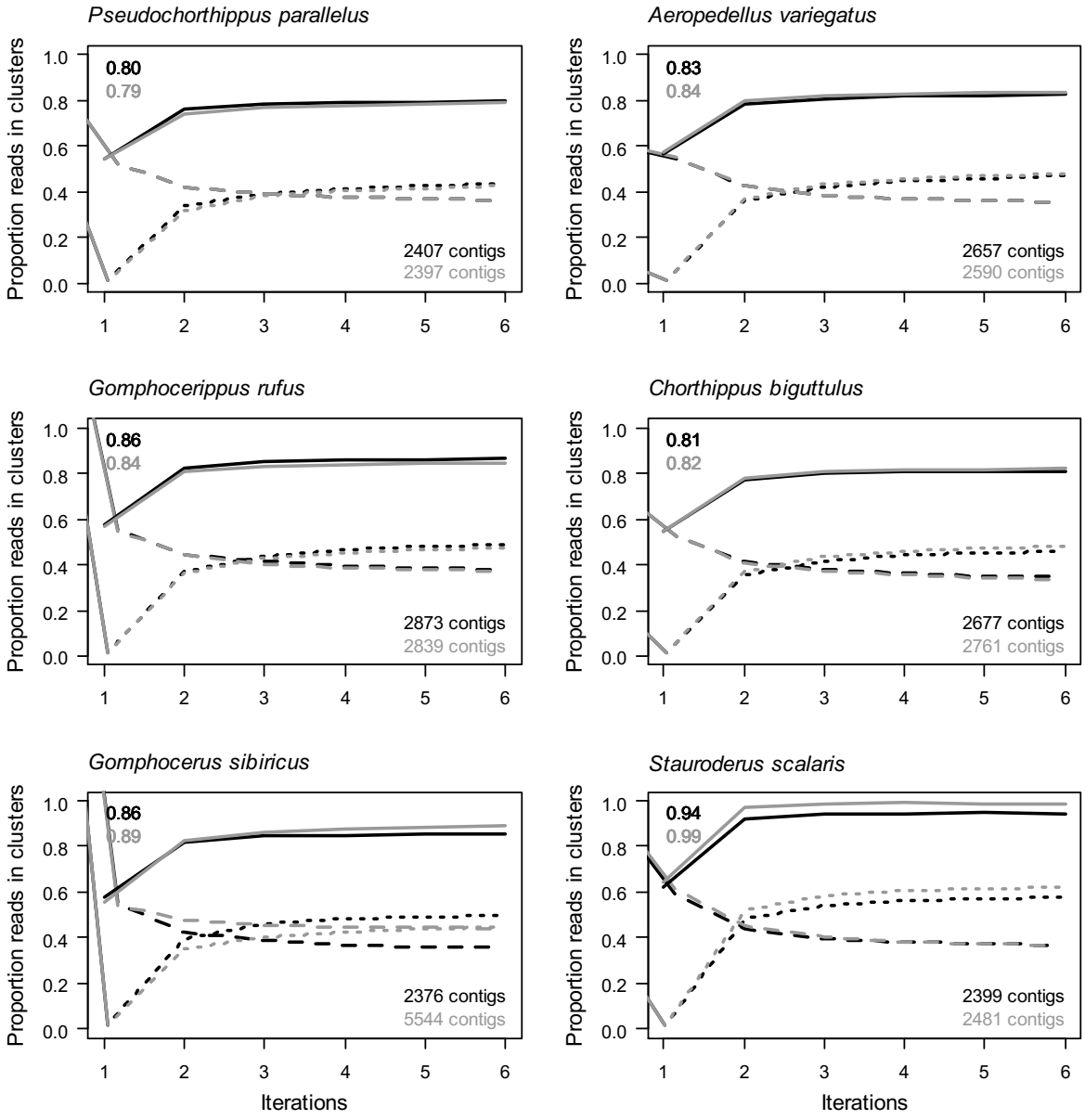


Figure S4: Distribution of repetitive DNA across repeat clusters annotated by RepeatMasker for six species of grasshopper. The data show results from a single female per species.

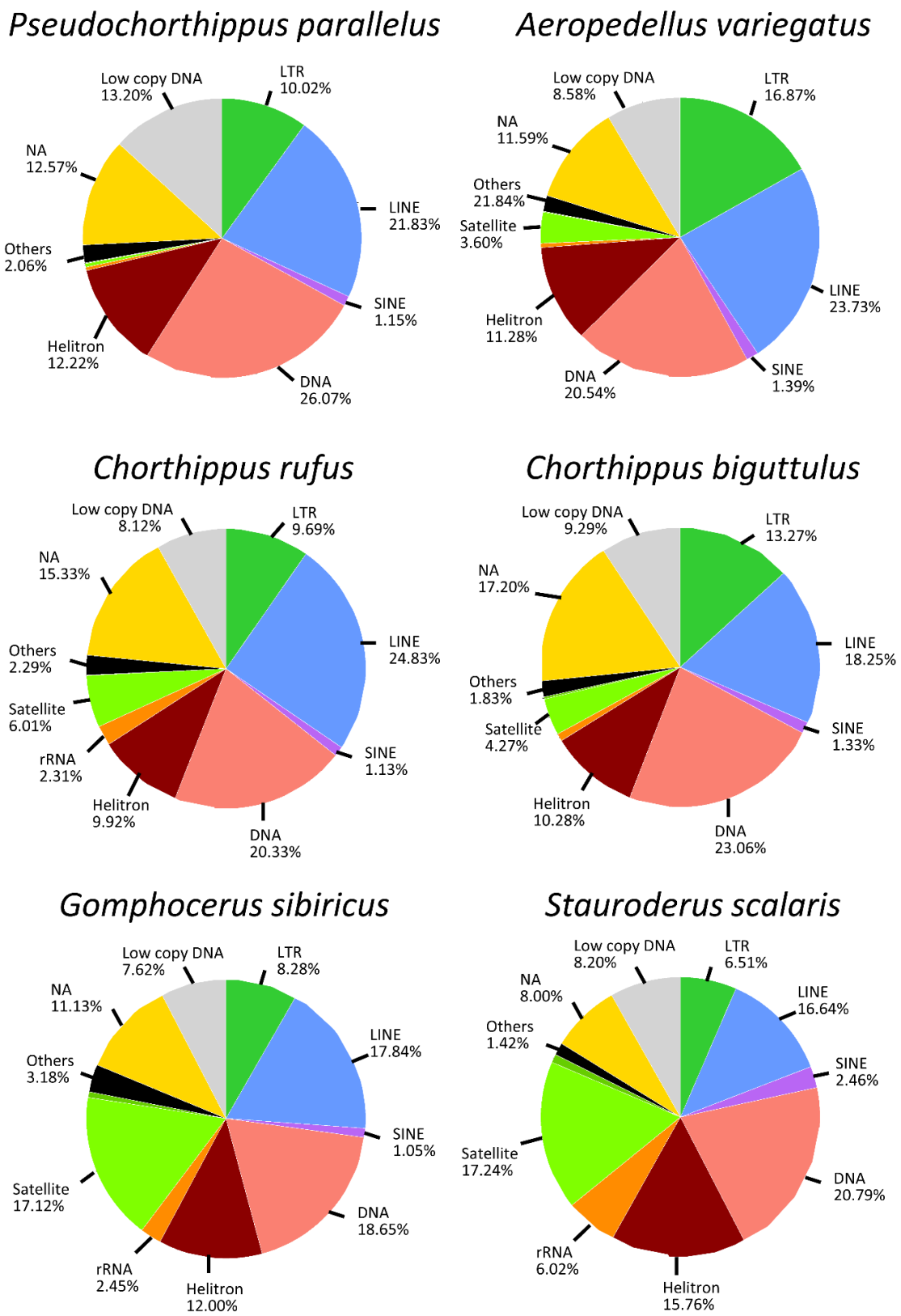


Figure S5: Distribution of repetitive DNA across repeat clusters annotated by RepeatMasker for six species of grasshopper. The data show results from a single male per species.

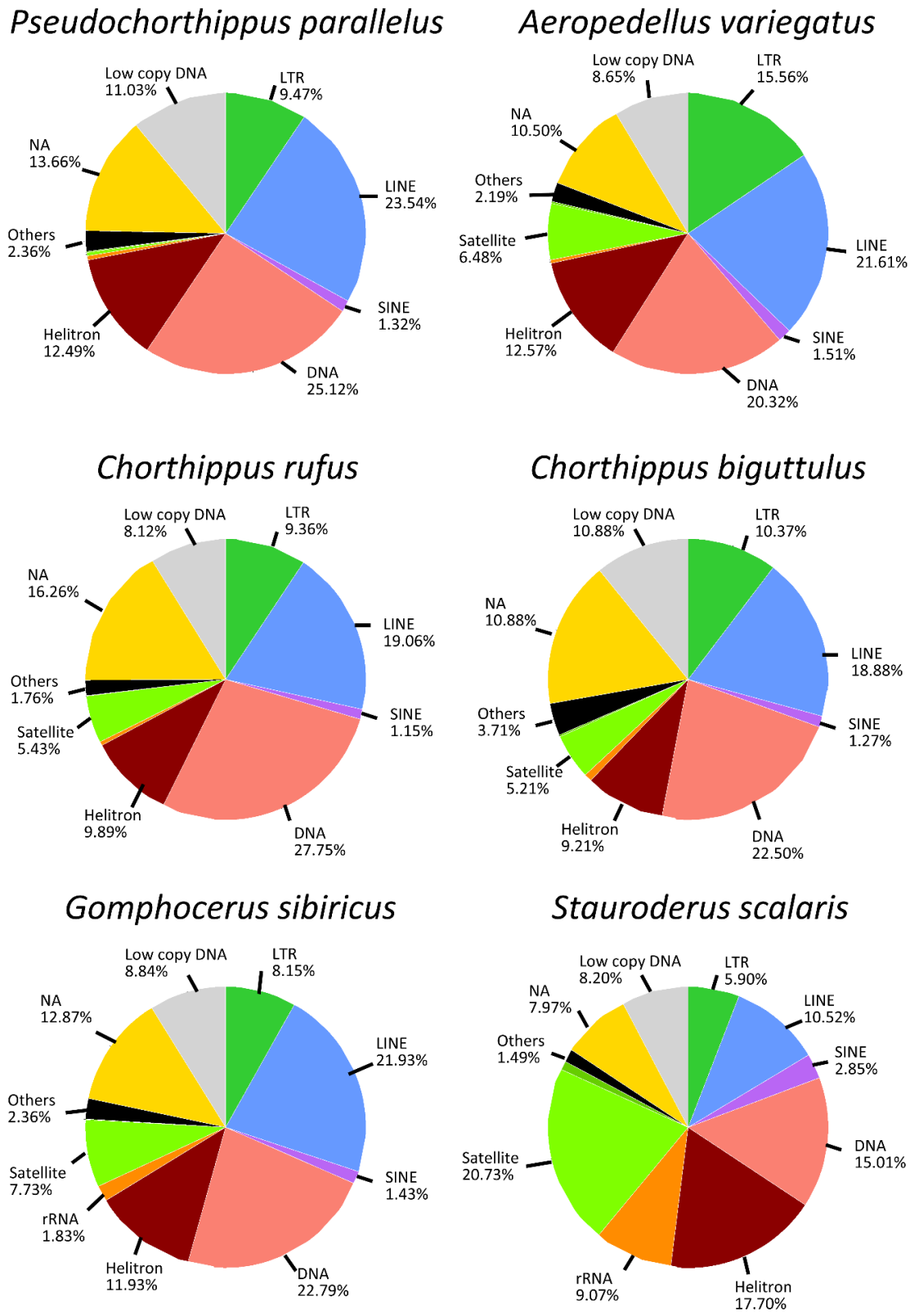


Figure S6: Distribution of de novo assembled repeat clusters for a *Locusta migratoria* female. Results are based on a single RepeatExplorer run. The vertical line shows the repeat content as estimated by RepeatExplorer based on this single run.

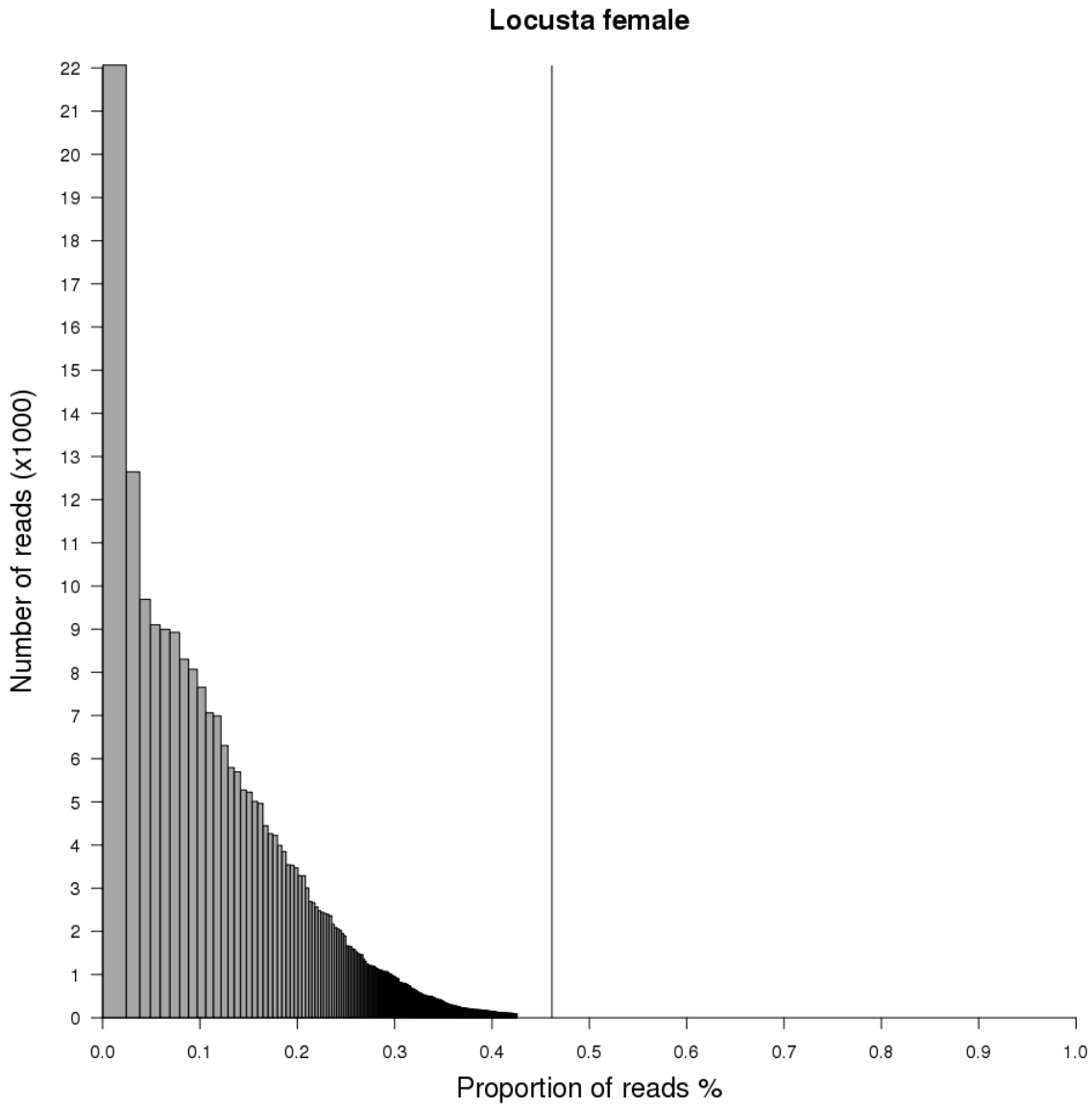


Figure S7: Divergence distribution within clusters of repetitive DNA in six species of grasshoppers across repeat-envirched datasets. The data show results from a single female per species.

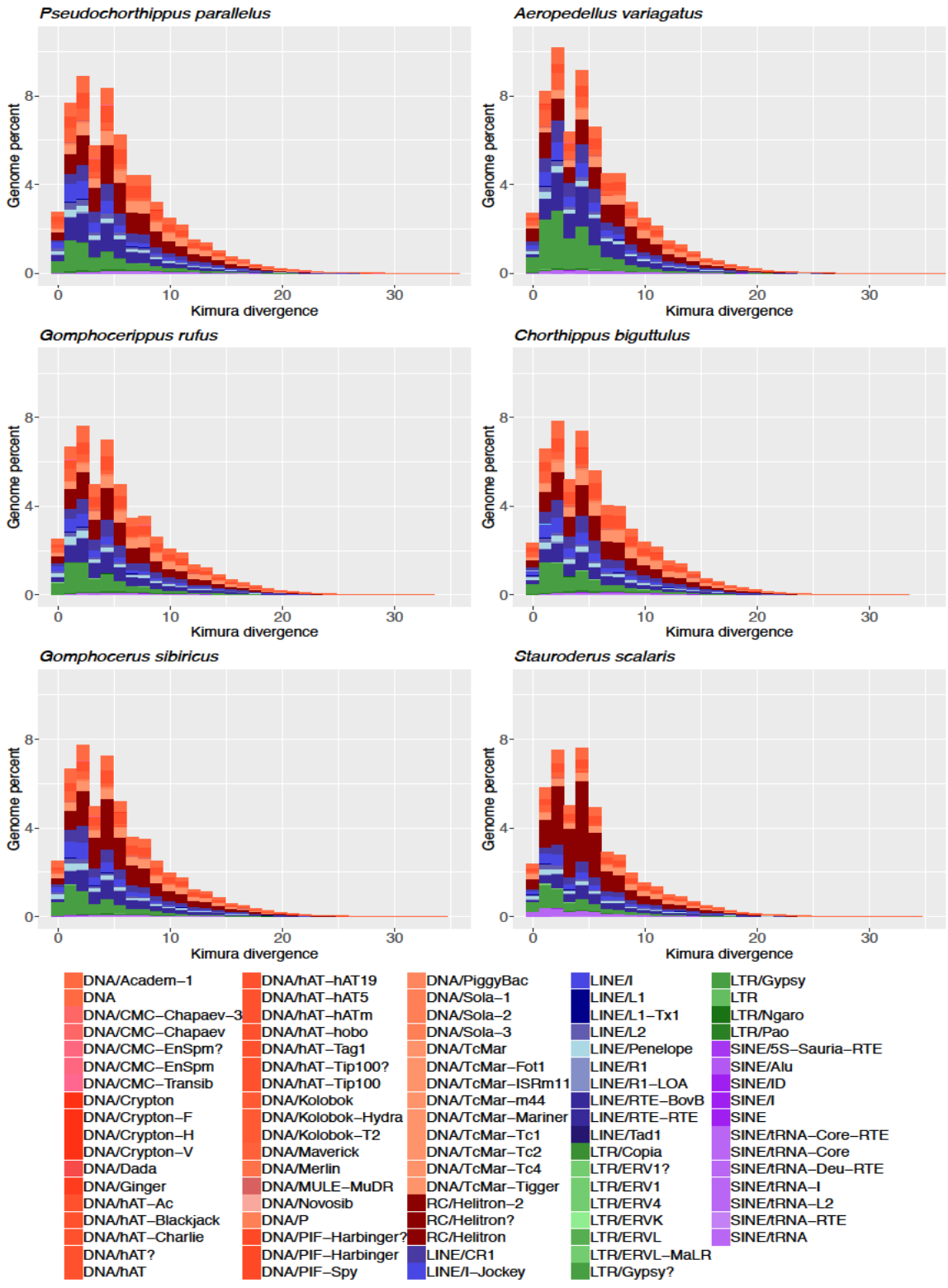


Figure S8: Divergence distribution within clusters of repetitive DNA in six species of grasshoppers across repeat-envirched datasets. The data show results from a single male per species.

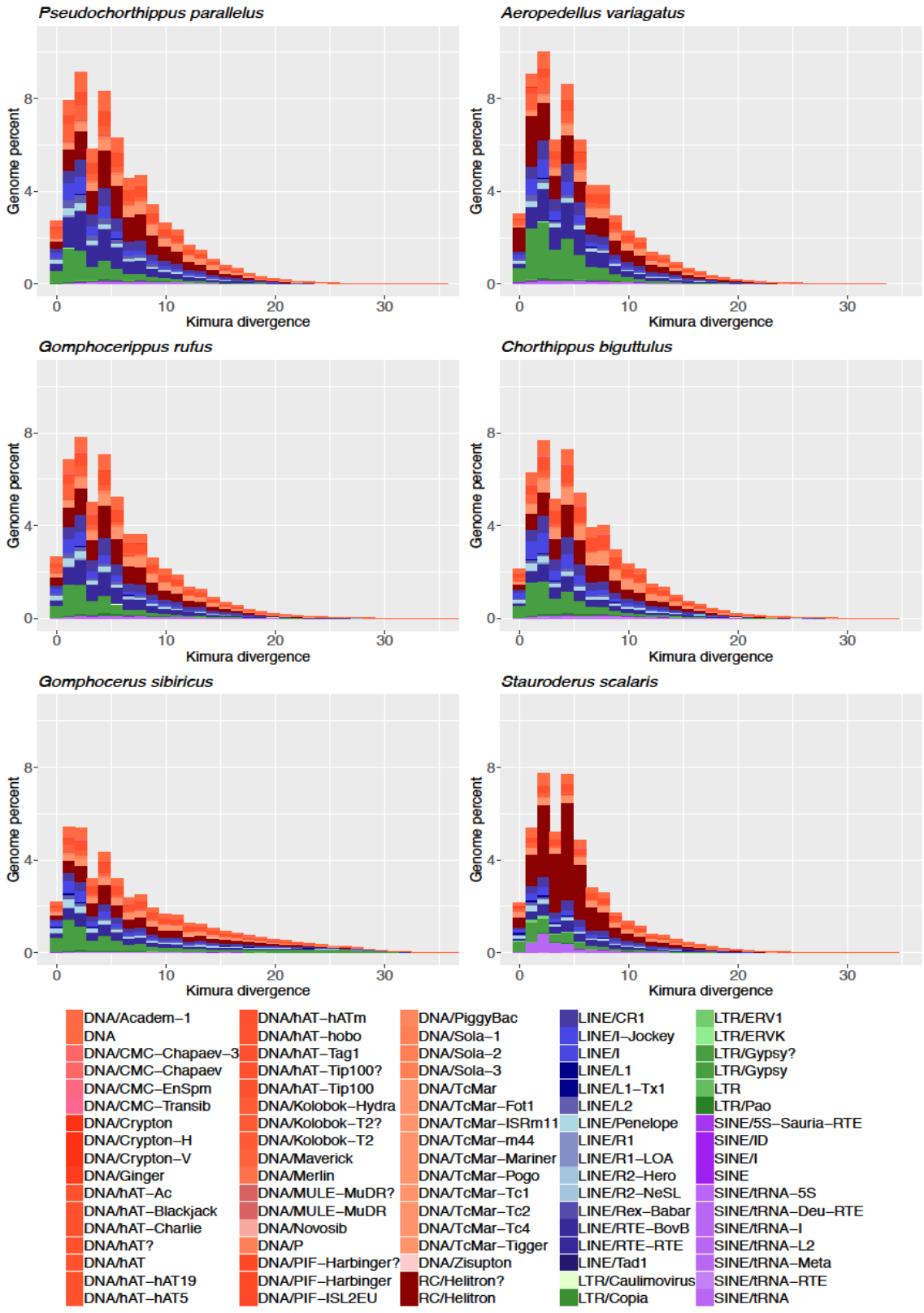


Figure S9: Correlation of repeat cluster size across the sexes. Results are shown across the first 10 repeat clusters with different clusters shown by different symbols. Each data point refers to a single cluster for a single species with six data points per cluster and thin lines showing the major axis regression line for each cluster. The dashed grey line shows the line of equal cluster size in females and males.

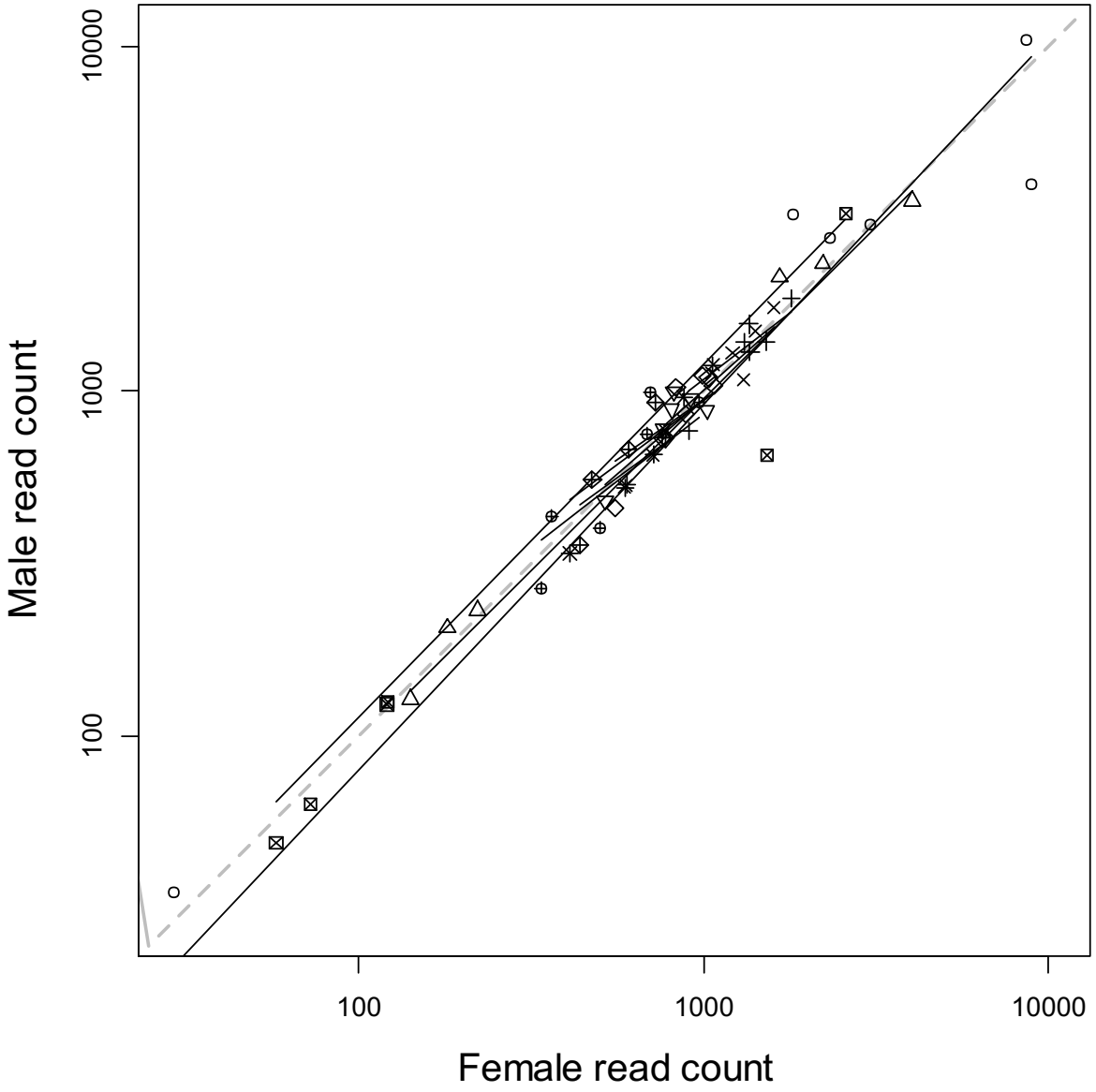


Figure S10: Sample specific phylogeny based on COI, COII and COIII mitrochondrial genes.

Figure S10: Sample specific phylogeny based on COI, COII and COIII mitrochondrial genes.

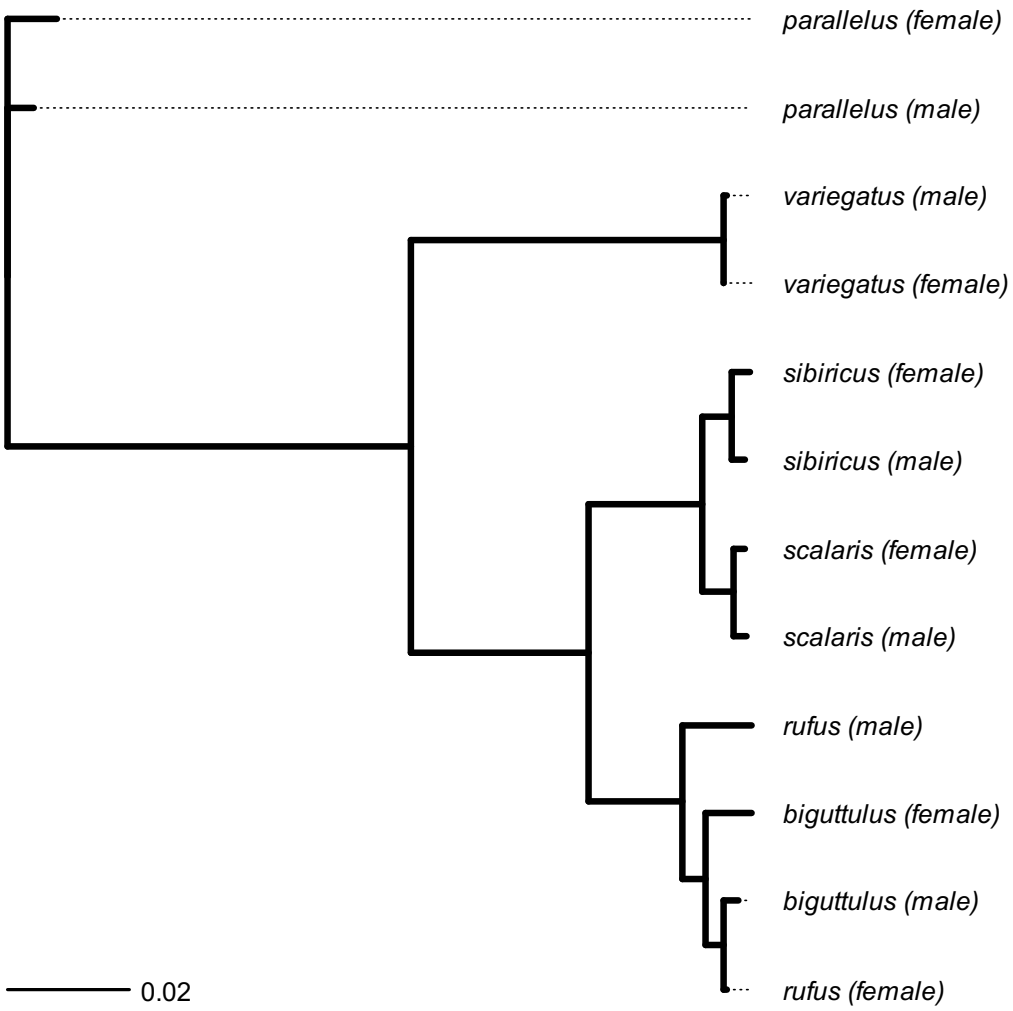


Figure S11: Sample-specific unrooted gene trees by mitochondrial gene locus.

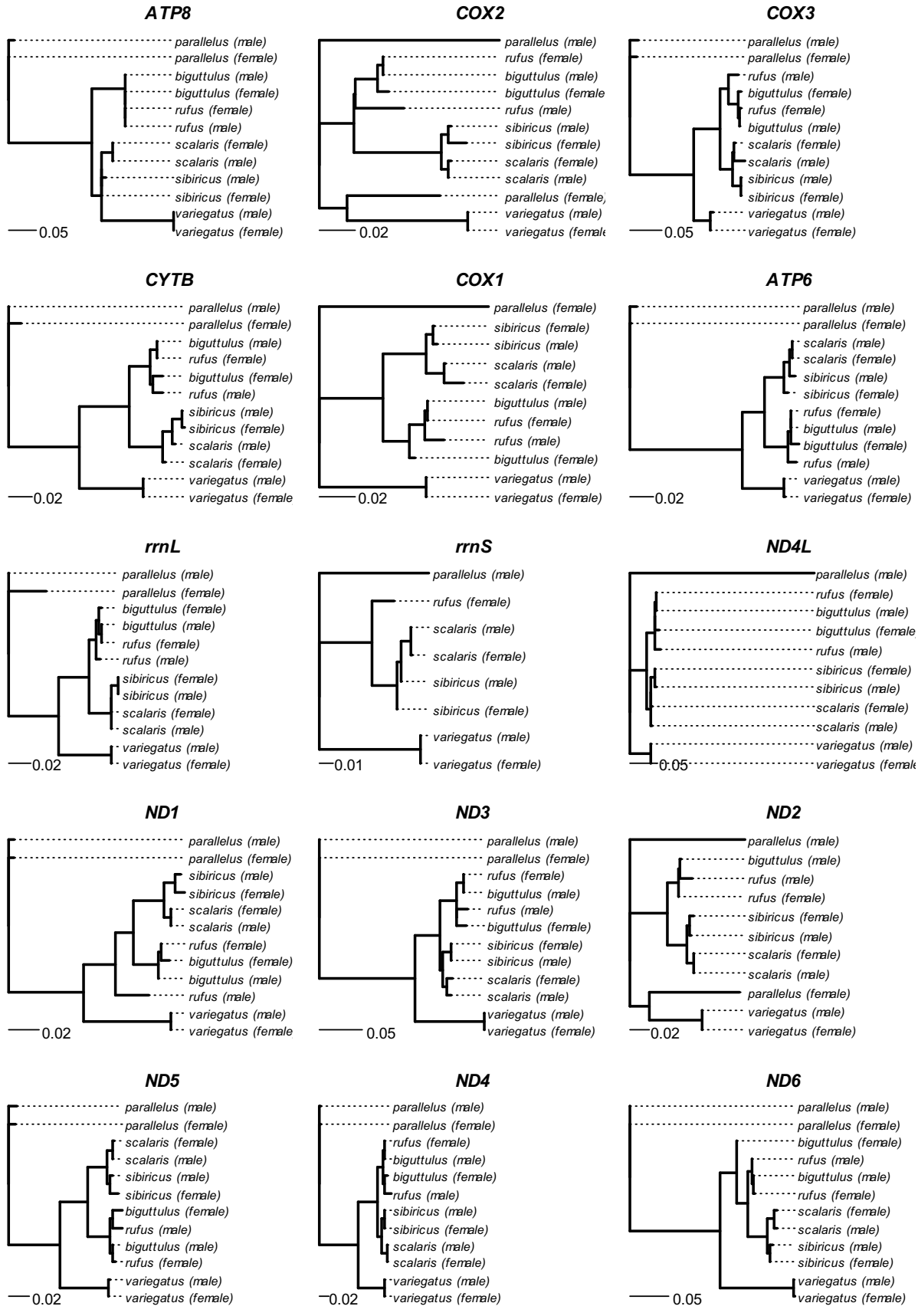


Figure S12: Repeat cluster size changes across the phylogeny. Grey dots show estimated abundance of reads (standardized to percentage of the largest node within a cluster). Edge width shows the proportional change (black = increases, red = decreases).

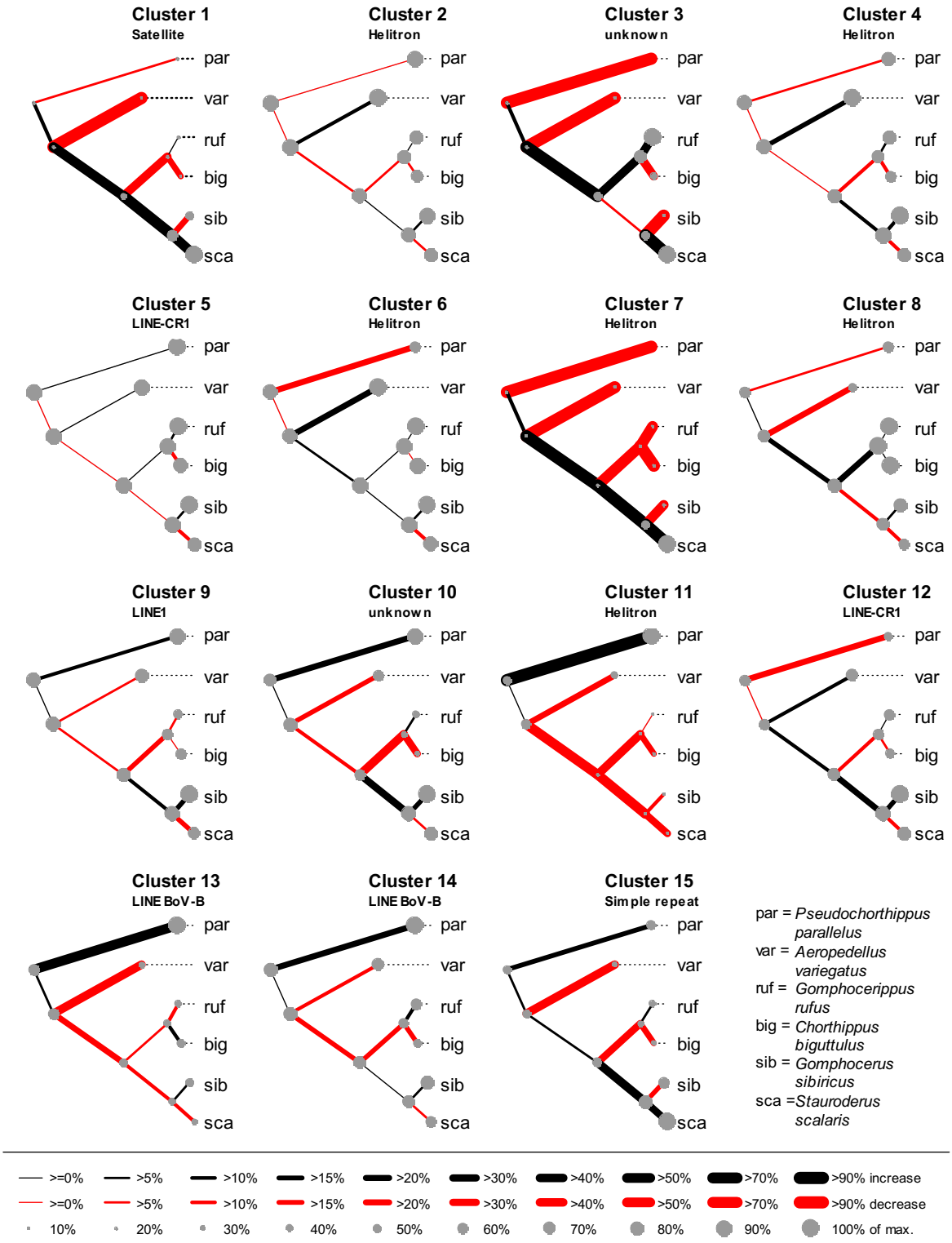


Figure S13**:** Cluster-paining approach to species-specific differences within cluster. The plot shows the four largest cluster with dots representing reads and read overlap by edges. The six different species are shown by different colours.

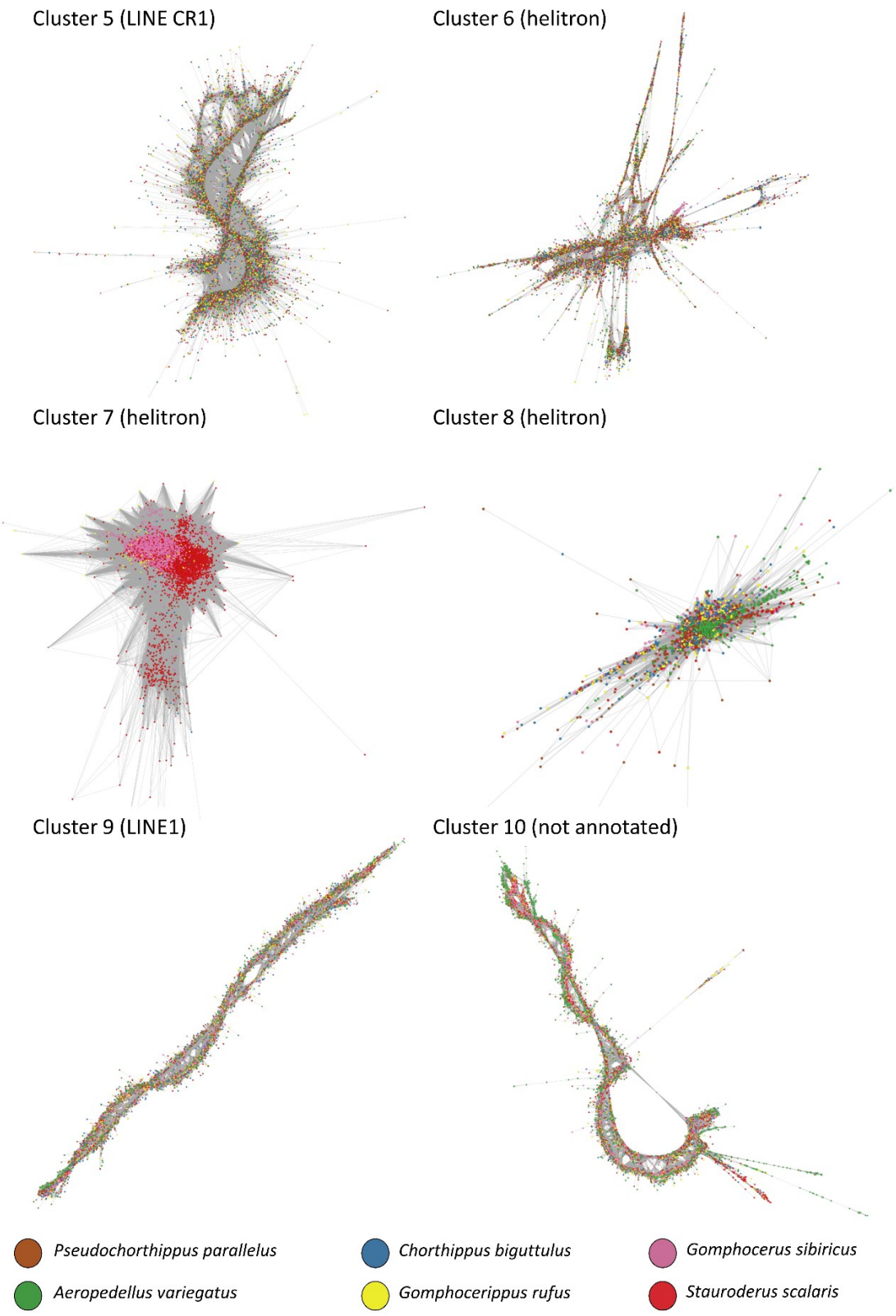


Figure S14**:** Calibration of flow cytometric signal intensity to genome sizes based on regression of published genome sizes (Table S3) and signal intensity. Published genome sizes are available for only four species (dom = *Acheta domesticus*, par = *Pseudochorthippus parallelus*, sib = *Gomphocerus sibiricus*, sca = *Stauroderus scalaris*, Table S3). Furthermore, we used the published genomes of *Chorthippus brunneus* as a substitute for the very closely related *Chorthippus biguttulus* (big/bru).

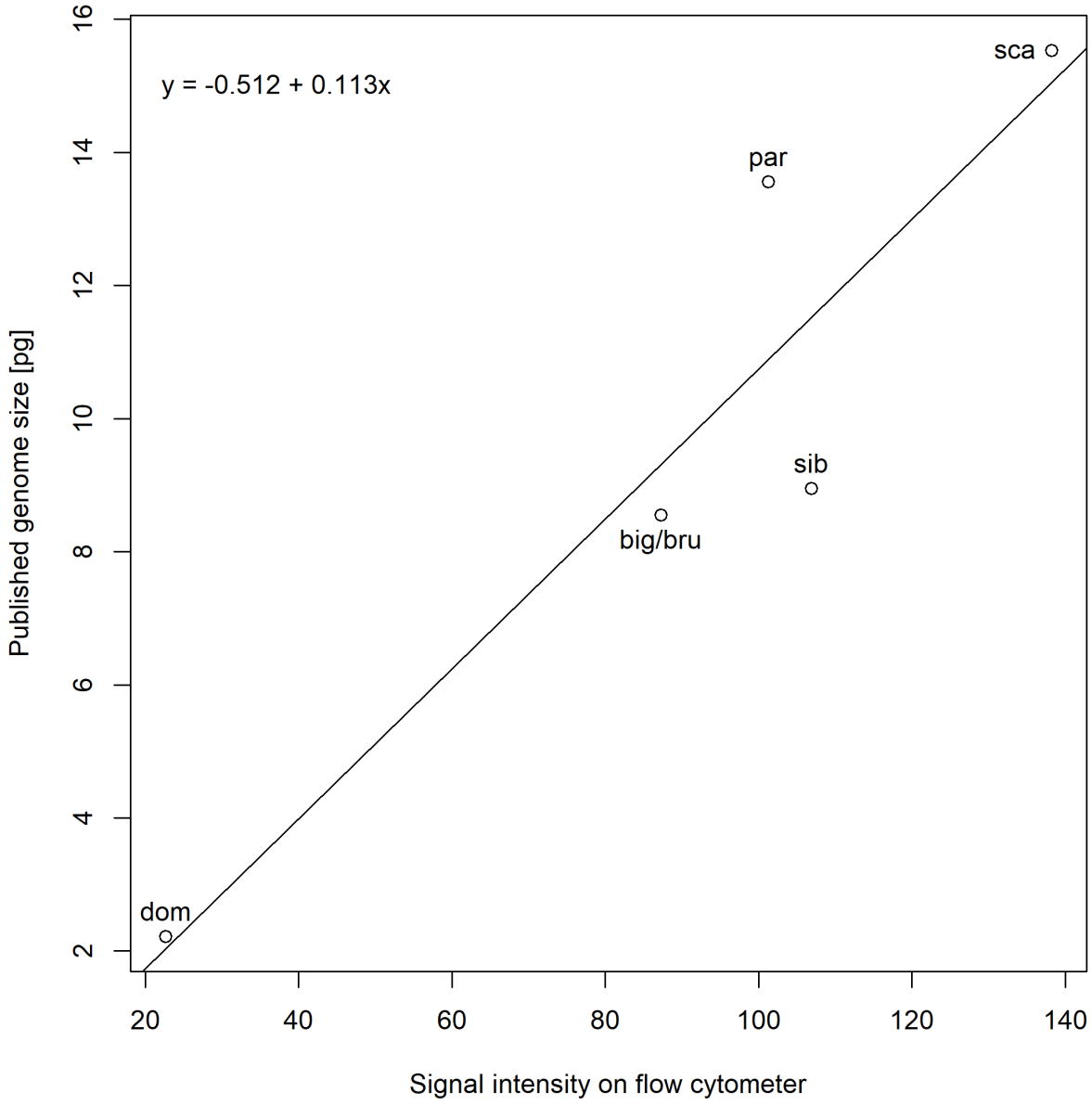


Table S1**:** Summary of sequencing output (after quality filtering) in seven species of Acridid grasshoppers. The *Locusta* dataset is a subset of the *Locusta* genome project sequencing runs (Wang *et al.*, 2014).

SPECIES	**SEX**	**# PE READS**	**# SE READS**	**TOTAL BP**	**COVERAGE**
Pseudochorthippus parallelus	Male	431,243	26,123,484	3,004,651,879	0.32x
Pseudochorthippus parallelus	Female	368,771	22,047,786	2,538,590,762	0.27x
Aeropedellus variegatus	Male	898,886	42,144,856	4,935,907,687	0.45x
Aeropedellus variegatus	Female	486,548	23,135,739	2,702,135,625	0.25x
Gomphocerippus rufus	Male	510,974	21,326,566	2,520,426,131	0.26x
Gomphocerippus rufus	Female	615,659	22,032,317	2,638,581,869	0.27x
Chorthippus biguttulus	Male	375,586	23,198,795	2,640,534,948	0.32x
Chorthippus biguttulus	Female	374,817	22,850,071	2,612,756,133	0.32x
Gomphocerus sibiricus	Male	463,973	19,980,621	2,376,345,690	0.23x
Gomphocerus sibiricus	Female	511,351	21,892,521	2,589,274,427	0.25x
Stauroderus scalaris	Male	375,939	28,034,183	2,448,458,508	0.18x
Stauroderus scalaris	Female	290,636	21,662,158	3,167,983,886	0.23x
Locusta migratoria	Female	8,420,257	0	7,482,931,189	1.28x

Table S2**:** Comparison of repeat content estimation by satMiner, RepeatExplorer and dnaPipeTE (correlations estimates satMiner-RepeatExplorer *r* = 0.56, satMiner-dnaPipeTE *r* = 0.93, RepeatExplorer-dnaPipeTE *r* = 0.72).

Species	Sex	satMiner	RepeatExplorer	dnaPipeTE
Pseudochorthippus parallelus	Male	0.79	0.68	0.61
Pseudochorthippus parallelus	Female	0.83	0.69	0.63
Aeropedellus variegatus	Male	0.84	0.73	0.64
Aeropedellus variegatus	Female	0.86	0.70	0.65
Gomphocerippus rufus	Male	0.84	0.69	0.64
Gomphocerippus rufus	Female	0.83	0.67	0.63
Chorthippus biguttulus	Male	0.82	0.67	0.62
Chorthippus biguttulus	Female	0.86	0.69	0.65
Gomphocerus sibiricus	Male	0.89	0.67	0.63
Gomphocerus sibiricus	Female	0.94	0.71	0.68
Stauroderus scalaris	Male	0.99	0.72	0.70
Stauroderus scalaris	Female	0.79	0.68	0.61

Table S3**:** Published genome size estimates as compiled in the Animal Genome Size Database (http://www.genomesize.com/). DNA content refers to haploid genome size. NA = Information not available.

SPECIES	DNA CONTENT [PG]	METHOD	CELL TYPE	SIZE STANDARD	ORIGIN	REFS
Pseudochorthippus parallelus	12.31	Fuelgen densitometry	Testes	*Locusta migratoria* (5.5 pg DNA)	UK?	John and Hewitt (1966)
Pseudochorthippus parallelus	13.36	Fuelgen densitometry	Testes	*Mus musculus* (3.3. pg DNA)	UK?	Wilmore and Brown (1975)
Pseudochorthippus parallelus	13.83	NA	NA	NA	NA	Petitpierre (1996)
Pseudochorthippus parallelus	14.72	Fuelgen densitometry	Testes	*Gallus domesticus* (1.25 pg DNA)	Spain	Belda *et al.* (1991)
Chorthippus brunneus	8.55	Fuelgen densitometry	Testes	*Locusta migratoria* (5.5 pg DNA)	UK?	John and Hewitt (1966)
Chorthippus brunneus	8.55	Fuelgen densitometry	Testes	*Locusta migratoria* (5.5 pg DNA)	UK?	Wilmore and Brown (1975)
Chorthippus brunneus	10.15	Fuelgen densitometry	Testes	*Allium cepa* (16.5 pg DNA)	Spain	Gosalvez *et al.* (1980)
Gomphocerus sibiricus	8.95	Fuelgen densitometry	Testes	*Allium cepa* (16.5 pg DNA)	Spain?	Gosalvez *et al.* (1980)
Stauroderus scalaris[1]	14.72	Fuelgen densitometry	Testes	*Gallus domesticus* (1.25 pg DNA)	Spain	Belda *et al.* (1991)
Stauroderus scalaris	16.34	NA	NA	NA	NA	Petitpierre (1996)

[1]Listed as Chorthippus scalaris.

Table S4**:** Sequence divergence within clusters summarized by repeat class and sample as estimated by dnaPipeTE.

Species	Sex	DNA transposons	Helitrons	LINE elements	LTR retro-transposons	SINE elements
Pseudochorthippus parallelus	Female	6.184	6.288	5.204	4.364	8.006
Pseudochorthippus parallelus	Male	6.354	6.463	5.255	4.313	7.995
Aeropedellus variegatus	Female	6.298	6.297	5.260	4.009	6.475
Aeropedellus variegatus	Male	6.293	5.215	5.270	4.071	6.635
Chorthippus biguttulus	Female	6.555	6.313	5.510	4.343	7.840
Chorthippus biguttulus	Male	6.681	6.313	5.522	4.298	7.654
Gomphocerippus rufus	Female	6.390	6.057	5.349	4.205	7.393
Gomphocerippus rufus	Male	6.222	6.167	5.372	4.123	7.518
Gomphocerus sibiricus	Female	6.207	5.872	5.219	4.101	7.193
Gomphocerus sibiricus	Male	7.616	7.996	7.043	7.865	9.245
Stauroderus scalaris	Female	6.168	4.944	5.510	4.012	4.735
Stauroderus scalaris	Male	6.171	4.738	5.387	4.050	4.291

Table S5**:** Matching of repeat clusters across independent runs here shown by example of runs 1 and 2. Rows show clusters as assigned in run 1 and columns show clusters as assigned in run 2. Numbers in cells show the number of reciprocal blast hits for contigs from each cluster in one run against all contigs from each clusters in the other run. In this particular example, the best-matching clusters for the first 20 clusters of each run are (Run 1-Run 2): 1-1, 2-3, 3-2, 4-4, 5-5, 6-6, 7-7, 8-8, 9-10, 10-13, 11-14, 12-12, 13-18, 14-15, 15-9, 16-17, 17-20, 18-16, 19-20, 20-11, 21-19. Among those, we consider the 17-20 match somewhat ambiguous.

		Cluster in Run 2																												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Cluster in run 1	**1**	9	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	2	·	·	13	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	1	·	·
	3	·	480	·	3	·	·	·	·	·	·	·	·	·	1	·	1	·	·	·	·	·	·	·	·	·	·	2	·	·	·
	4	·	5	·	229	·	1	·	·	·	·	·	1	·	·	1	·	·	1	·	·	·	·	·	·	·	·	·	·	·	·
	5	·	·	·	·	50	·	·	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	6	·	1	·	·	·	109	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	7	·	·	·	·	·	·	11	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	8	·	·	·	1	·	·	·	22	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	9	·	·	·	·	·	·	·	·	·	29	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	10	·	·	·	·	·	·	·	·	·	·	·	·	66	·	·	·	·	1	·	·	5	·	·	·	·	·	·	·	·	·
	11	·	·	·	·	·	·	·	·	·	·	3	·	·	35	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	12	·	·	·	·	·	·	·	·	·	·	·	42	·	·	·	·	·	·	·	·	2	·	·	·	2	·	·	·	·	·
	13	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	26	2	1	·	·	·	·	·	·	·	·	·	·
	14	·	·	·	·	·	·	·	·	·	·	3	·	·	·	37	·	·	·	·	13	·	·	·	·	·	·	·	·	2	·
	15	·	·	·	·	·	·	·	·	221	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·
	16	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	203	·	·	·	·	·	·	·	·	·	·	·	·	·
	17	·	·	·	·	·	·	·	·	·	·	·	·	5	·	·	·	·	1	·	8	·	·	·	·	·	·	·	·	·	·
	18	·	1	·	5	·	·	·	·	·	·	·	·	·	·	·	119	·	·	·	·	·	·	·	·	1	·	·	·	·	4
	19	·	·	·	·	·	·	·	·	·	·	1	·	·	·	3	·	·	·	·	52	·	·	·	·	·	·	·	·	7	·
	20	·	·	·	·	·	·	·	·	·	·	154	·	·	·	·	·	·	·	·	1	2	·	·	·	·	·	·	·	·	·
	21	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	17	·	·	·	·	·	·	·	·	·	·	·
	22	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	9	·	·	·	·	·	·	·	·
	23	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	9	·	·	·	·	·	·
	24	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	6	·	·	·	·	·	·	·
	25	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	18	·	·	·	·	·
	26	·	·	·	·	·	·	·	·	·	·	12	·	·	·	·	·	·	·	·	·	141	·	·	·	·	·	·	·	·	·
	27	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	30	·	·	·	·
	28	·	30	·	1	·	·	·	1	·	·	1	1	·	1	·	2	·	·	·	·	·	·	·	·	·	·	3	·	·	2
	29	·	·	·	·	·	·	·	·	·	·	1	·	·	·	·	·	·	·	·	13	·	·	·	·	·	·	·	·	23	·
	30	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	5	·	·

Table S6**:** Rank order in which the top 15 most abundant clusters appear in each of the 20 independent *de novo* repeat assembly runs on data pooled from six species of Gomphocerine grasshoppers.

	RUN																			
CLUSTER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	3	3	3	2	3	2	3	2	2	3	2	2	2	2	3	2	2	2	2	2
3	2	2	2	3	2	3	2	3	3	2	3	3	3	3	2	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	6	5	6	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	7	6	5	6	6	6	6	6	6
7	7	7	7	7	7	8	7	7	7	7	7	8	7	7	7	7	7	7	8	7
8	8	8	8	8	8	9	8	8	8	8	8	9	8	8	8	8	8	8	9	8
9	9	10	9	9	10	10	9	9	9	10	9	11	11	11	9	9	9	9	10	9
10	10	11	11	10	11	11	10	10	10	12	11	12	13	12	11	12	10	11	11	11
11	19	12	13	11	9	19	12	11	13	9	10	10	12	10	10	11	19	10	12	19
12	12	14	12	13	12	13	13	12	11	14	12	13	14	13	12	13	11	12	13	12
13	11	13	14	12	13	12	11	13	12	15	13	14	15	14	13	14	12	13	14	13
14	14	15	15	14	14	14	15	14	14	16	15	16	16	15	14	15	13	15	7	14
15	13	16	16	16	15	16	14	15	20	18	14	17	19	17	16	16	16	17	16	15

Table S7**:** Annotation of repeat clusters using RepeatMasker with the Metazoan database of repeats. The table shows all cases with more than 1% hits.

CLUSTER	CLASS	FAMILY	# HITS	HITS [%]
1	Satellite	Satellite	108,700	92.3%
2	DNA	DNA	13,135	20.7%
	RC	Helitron?	7,849	15.1%
	RC	Helitron	6,216	8.7%
	DNA	hAT-Charlie	2,281	3.2%
	Unknown	Unknown	1,291	1.6%
3	-	-	-	all <1.0%
4	RC	Helitron	28,700	72.0%
	Unknown	Unknown	704	1.5%
	DNA	hAT-Ac	888	2.0%
	Unknown	Unknown	704	1.5%
5	LINE	CR1	21,554	95.3%
6	RC	Helitron	14,331	54.7%
	DNA	hAT-Tip100	2,404	9.2%
	LINE	CR1	1,413	3.4%
	DNA	hAT-Blackjack	491	1.4%
7	RC	Helitron	5,130	27.7%
8	RC	Helitron	10,494	46.7%
	DNA	TcMar-Tc1	5,405	15.5%
9	LINE	I	15,349	96.7%
10	tRNA	tRNA	1,088	2.37%

11	RC	Helitorn	12,521	55.6%
12	LINE	CR1	13,919	92.8%
13	LINE	BovB	13,572	95.7%
14	LINE	BovB	10,994	75.7%
15	-	-	-	all <1.0%

General Discussion

Main Findings

The dramatic decline of high throughput sequencing and computational costs in the last few years have enabled us to explore many less studied organisms which was previously not possible (Wetterstrand). Furthermore, recent advances in bioinformatics methods has also enabled us to tackle significant challenges in terms of informational, genomic and computational complexity. This has allowed us to explore species with enormous and complex genomes chalk full of repetitive elements.

In this thesis, manuscript I demonstrated the utility of high-throughput sequencing and prior bioinformatics analysis to reduce the costs and improve the efficiency of traditional genetic approaches such as using microsatellite markers for exploring species with complex genomes. The repeat landscape and proportion of repetitive DNA in the genome, highly informative statistics, were also estimated along with our main goals of developing viable microsatellite marker. This method will definitely allow us to investigate many other species which have been difficult to work with due to genomic complexity.

Manuscript II described the assembly and analysis of transcriptome of species (*Gomphocerus sibiricus*) with a large and complex genome. Here, I was also able to reconstruct the mitochondrial genome which suggested that another closely related species (*Gomphocerus licenti*) was less divergent than another population of the same species. I also provided additional evidence for the presence of the endo-parasite *Wolbachia*. Furthermore, this also paves the way for future expression analysis studies such as green brown polymorphism, development of club ornaments in *Gomphocerus sibiricus*.

Manuscript III explored the genome size through a comparative genomics approach. We analyzed the repetitive DNA of 6 species, which also included *Stauroderus scalaris*, an insect with an estimated genome size of approximately 14 GB and estimated repeat content over 95%. Overall, we observed positive correlation between genome size and repeat content. Here, the repetitive sequence content estimate of *Locusta migratoria* was in line with the estimate from the assembled genome of the species. Furthermore, my main findings suggested that expansion of satellite DNA in *Gomphocerus sibiricus* and *Stauroderus scalaris*,

and this is rather a consequence of genome size expansion than a cause. I also suggest that expansion in satellite DNA elements is to stabilize chromosome structure. Overall, genome size expansion may occur in a complex multi-step process where various phases are governed by varied processes.

Methodological Advances and Future Directions

This dissertation is a testament to the potential possibilities of optimizing and reducing costs of investigating biological assays by employing cutting-edge bioinformatics methods in the design process.

In manuscript I, I developed a new method to maximize the discovery of polymorphic microsatellites from complex genomes, thereby reducing overall costs of using such markers in classical genetic studies. This methodological advancement has already yielded benefits in studies involving the bow winged grasshopper, *Chorthippus biguttulus*, where it was used for parentage analysis in order to study cryptic female choice. This study was carried out by a PhD student, Michael Haneke-Reinders at Bielefeld University. This experimentally demonstrated improvement in efficiency will certainly be appreciated by other future studies on species with complex genomes.

In manuscript II, I assembled and annotated a high-quality reference transcriptome. In most modern reference transcriptomes are usually the result of the consensus of many assemblies of multiple *k-mer* lengths. However, as more assemblies are used to create consensus contigs, the risk of assembling chimeric contigs also significantly increases. Here, I screened individual contig assembly metrics to lower this risk considerably. Thereby improving upon already available reference transcriptome assembly protocols. *Gomphocerus sibiricus* is now poised to become a model system to model study system for behaviour, quantitative genetics, green-brown dimorphism, sexual selection and perhaps orthopteran genomics.

In manuscript III, I surveyed and compared complex insect genomes using low coverage (<0.45X) sequencing data. The *de novo* discovery of TEs and other repetitive elements are incredibly compute and memory intensive pipelines which require considerable resources. Furthermore, the pragmatic upper limit of data for the RepeatExplorer pipeline is about 400

million bases (Novák,P. 2017, personal communication). This scales considerable with repetitive content especially during the graph based clustering phase and the all-to-all similarity search steps of the pipeline. Although RepeatExplorer automatically estimates an appropriate sample size, the run time of the pipeline more than 32,000 CPU hours. Furthermore, non-repetitive read sequences are also included in the sequence input of typical runs (in our case about 30-40%), which, however, are mostly uninformative to the pipeline, and therefore contribute to the inefficiency of the pipeline. In manuscript III, I describe a method, which yields repeat enriched read-sets which have an average repetitive content of over 91%. Furthermore, my method of running multiple sub-samples followed by re-ordering of clusters, instead of a single large run allows us to further speedup the run times by leveraging the multiple compute nodes on various subsets of repeat enriched reads. In the future, it will be possible to run high coverage data-sets in this fashion to fully unravel the complexity of the TE landscapes of even the largest and most complex genomes.

In the future it we may also extend this approach in other directions. In essence, the repeat enrichment step partitions the input read data into reads which are likely to contain repetitive sequences and reads which are unlikely to contain repetitive reads. However, in our study we did not fully utilize the information from the reads unlikely to contain repetitive sequences. A “low-hanging fruit” would be to assemble the non-repetitive faction of the genome from these reads as they would represent the faction of the genome most computationally easy. As most genome assembly pipelines extensively exploit DGBs, repetitive sequences pose a significant challenge due to their repetitive nature, which in many instances causes the constructed graph to become extremely large and perhaps unresolvable. Assembling the non-repetitive reads would probably yield a partial genome assembly with perhaps many fully intact genes, albeit highly fragmented and with significant missing portions.

References

Alfsnes, K., Leinaas, H. P., & Hessen, D. O. (2017). Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. *Ecology and Evolution, 7*, 5939-5947.

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., et al. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources, 20*, 892-905.

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

Andersen, A. N., Ludwig, J. A., Lowe, L. M., & Rentz, D. C. F. (2001). Grasshopper biodiversity and bioindicators in Australian tropical savannas: Responses to disturbance in Kakadu National Park. *Austral Ecology, 26*, 213-222.

Anstey, M. L., Rogers, S. M., Ott, S. R., Burrows, M., & Simpson, S. J. (2009). Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *Science, 323*, 627.

Bakkali, M., & Martín-Blázquez, R. (2018). RNA-Seq reveals large quantitative differences between the transcriptomes of outbreak and non-outbreak locusts. *Scientific Reports, 8*, 9207.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA, 6*, 11.

Belda, J. E., Cabrero, J., Camacho, J. P. M., & Rufas, J. S. (1991). Role of C-heterochromatin in variation of nuclear DNA amount in the genus *Chorthippus* (Orthoptera, Acrididae). *Cytobios, 67*, 13-21.

Bellmann, H., & Luquet, G.-C. (2009). *Guide des sauterelles, grillons et criquets d'Europe occidentale*. Paris: Delachaux et Niestlé.

Berdan, E. L., Finck, J., Johnston, P. R., Waurick, I., Mazzoni, C. J., et al. (2017). Transcriptome profiling of ontogeny in the acridid grasshopper *Chorthippus biguttulus*. *PLoS ONE, 12*, e0177367.

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

Branson, D. H., Joern, A., & Sword, G. A. (2006). Sustainable management of insect herbivores in grassland ecosystems: new perspectives in grasshopper control. *BioScience, 56*, 743-755.

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., et al. (2017). A Tissue-Mapped Axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell Reports, 18*, 762-776.

Burri, R. (2017). Linked selection, demography and the evolution of correlated genomic landscapes in birds and beyond. *Molecular Ecology, 26*, 3853-3856.

Caballero, J., Smit, A. F. A., Hood, L., & Glusman, G. (2014). Realistic artificial DNA sequences as negative controls for computational genomics. *Nucleic Acids Research, 42*, e99.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., & Olmo, E. (2016). Transposons, genome size, and evolutionary insights in animals. *Cytogenetic and Genome Research, 147*, 217-239.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science, 34*, 247-278.

Cerveau, N., & Jackson, D. J. (2016). Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC Bioinformatics, 17*, 525.
Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature, 371*, 215-220.
Chu, C., Nielsen, R., & Wu, Y. (2016). REPdenovo: Inferring *de novo* repeat motifs from short sequence reads. *PLoS ONE, 11*, e0150719.
Cigliano, M. M., H. Braun, D.C. Eades, & Otte, D. (2018). Orthoptera Species File. (Web Database), http://Orthoptera.SpeciesFile.org, Retrieved 01-11-2018
Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13.
Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLoS Biology, 13*, e1002112.
da Silva, A. F., Dezordi, F. Z., Loreto, E. L. S., & Wallau, G. L. (2018). *Drosophila* parasitoid wasps bears a distinct DNA transposon profile. *Mobile DNA, 9*, 23.
Devkota, B., & Schmidt, G. H. (2000). Accumulation of heavy metals in food plants and grasshoppers from the Taigetos Mountains, Greece. *Agriculture, Ecosystems & Environment, 78*, 85-91.
Dieker, P., Beckmann, L., Teckentrup, J., & Schielzeth, H. (2018). Spatial analyses of two color polymorphisms in an alpine grasshopper reveal a role of small-scale heterogeneity. *Ecology and Evolution, 20*, 235-212.
Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature, 284*, 601-603.
Dumas, P., Tetreau, G., & Petit, D. (2010). Why certain male grasshoppers have clubbed antennae? *Comptes Rendus Biologies, 333*, 429-437.
Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics, 17*, 422-433.
Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., et al. (2012). The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature, 491*, 756-760.
Elliott, T. A., & Gregory, T. R. (2015). Do larger genomes contain more diverse transposable elements? *BMC Evolutionary Biology, 15*, 69.
Feliciello, I., Akrap, I., Brajković, J., Zlatar, I., & Ugarković, Đ. (2014). Satellite DNA as a driver of population divergence in the red flour beetle *Tribolium castaneum*. *Genome Biology and Evolution, 7*, 228-239.
Garner, T. W. (2002). Genome size and microsatellites: the effect of nuclear size on amplification potential. *Genome, 45*, 212-215.
Garrido-Ramos, M. A. (2017). Satellite DNA: an evolving topic. *Genes, 8*, 230.
Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics, 19*, 688-704.
Gordon, A., & Hannon, G. J. (2010). FASTX-toolkit: FASTQ/A short-reads preprocessing tools. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/index.html
Gosalvez, J., & López-Fernandez, C. (1981). Extra heterochromatin in natural populations of *Gomphocerus sibiricus* (Orthoptera: Acrididae). *Genetica, 56*, 197-204.
Gosalvez, J., López-Fernandez, C., & Esponda, P. (1980). Variability of the DNA Content in Five Orthopteran Species. *Caryologia, 33*, 275-281.
Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., et al. (2015a). *De novo* assembly and annotation of the asian tiger mosquito (*Aedes albopictus*) repeatome

with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution, 7*, 1192-1205.
Goubert, C., Modolo, L., Vieira, C., ValienteMoro, C., Mavingui, P., et al. (2015b). De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution, 7*, 1192-1205.
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology, 29*, 644-652.
Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews, 76*, 65-101.
Gregory, T. R. (2005). Genome size evolution in animals. In T. R. Gregory (Ed.), *The Evolution of the Genome* (pp. 3-87). Burlington: Academic Press.
Gregory, T. R. (2018). Animal Genome Size Database. http://www.genomesize.com/, Retrieved 09-05-2018
Guignard, M. S., Nichols, R. A., Knell, R. J., Macdonald, A., Romila, C.-A., et al. (2016). Genome size and ploidy influence angiosperm species; biomass under nitrogen and phosphorus limitation. *New Phytologist, 210*, 1195-1206.
Guo, R., Li, Y.-R., He, S., Ou-Yang, L., Sun, Y., et al. (2017). RepLong: *de novo* repeat identification using long read sequencing data. *Bioinformatics, 34*, 1099-1107.
Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols, 8*, 1494-1512.
Hare, E. E., & Johnston, J. S. (2011). Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods in Molecular Biology, 772*, 3-12.
Hartl, D. L. (2000). Molecular melodies in high and low C. *Nature Reviews Genetics, 1*, 145-149.
Hartl, D. L., & Clark, A. G. (2007). *Principles of Population Genetics* (4th ed.): Sunderland: Sinauer and Associates.
Hawlitschek, O., Morinière, J., Lehmann, G. U. C., Lehmann, A. W., Kropf, M., et al. (2017). DNA barcoding of crickets, katydids and grasshoppers (Orthoptera) from Central Europe with focus on Austria, Germany and Switzerland. *Molecular Ecology Resources, 17*, 1037-1053.
Heifetz, Y., Voet, H., & Applebaum, S. W. (1996). Factors affecting behavioral phase transition in the desert locust, *Schistocerca gregaria* (Forskål) (Orthoptera: Acrididae). *Journal of Chemical Ecology, 22*, 1717-1734.
Hollister, J. D., & Gaut, B. S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research, 19*, 1419-1428.
Jetybayev, I., Bugrov, A., Dzuybenko, V., & Rubtsov, N. (2018). B chromosomes in grasshoppers: different origins and pathways to the modern Bs. *Genes, 9*, 509-519.
Jiang, F., Yang, M., Guo, W., Wang, X., & Kang, L. (2012). Large-scale transcriptome analysis of retroelements in the migratory locust, *Locusta migratoria*. *PLoS ONE, 7*, e40532.
John, B., & Hewitt, G. M. (1966). Karyotype stability and DNA variability in the Acrididae. *Chromosoma, 20*, 155-172.
Kapitonov, V. V., & Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics, 23*, 521-529.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences USA, 114*, E1460-E1469.
Katoh, K., & Standley, D. M. (2013). MAFFT: Multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 30*, 772-780.
Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature, 217*, 624-626.
Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
Koch, P., Platzer, M., & Downie, B. R. (2014). RepARK—*de novo* creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research, 42*, e80-e80.
Köhler, G., Samietz, J., & Schielzeth, H. (2017). Morphological and colour morph clines along an altitudinal gradient in the meadow grasshopper *Pseudochorthippus parallelus*. *PLoS ONE, 12*, e0189815.
Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2016). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution, 34*, 772-773.
Laws, A. N., Prather, C. M., Branson, D. H., & Pennings, S. C. (2018). Effects of grasshoppers on prairies: Herbivore composition matters more than richness in three grassland ecosystems. *Journal of Animal Ecology, 87*, 1727-1737.
Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., et al. (2017). Less effective selection leads to larger genomes. *Genome Research, 27*, 1016-1028.
Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.
Li, R., Wang, Y., Shu, X., Meng, L., & Li, B. (2020). Complete mitochondrial genomes of three *Oxya* grasshoppers (Orthoptera) and their implications for phylogenetic reconstruction. *Genomics, 112*, 289-296.
Lower, S. S., Johnston, J. S., Stanger-Hall, K. F., Hjelmen, C. E., Hanrahan, S. J., et al. (2017). Genome size in North American fireflies: substantial variation likely driven by neutral processes. *Genome Biology and Evolution, 9*, 1499-1512.
Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience, 4*, 18.
Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA.: Sinauer Associates.
Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science, 302*, 1401-1404.
Maumus, F., Fiston-Lavier, A.-S., & Quesneville, H. (2015). Impact of transposable elements on insect genomes and biology. *Current Opinion in Insect Science, 7*, 30-36.
Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science, 346*, 763-767.
Novák, P., Neumann, P., & Macas, J. (2010). Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics, 11*, 378.
Novák, P., Neumann, P., Pech, J., Steinhaisl, J., & Macas, J. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics, 29*, 792-793.
Nowoshilow, S., Schloissnig, S., Fei, J.-F., Dahl, A., Pang, A. W. C., et al. (2018). The axolotl genome and the evolution of key tissue formation regulators. *Nature, 554*, 50-55.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature, 497*, 579-584.
O'Neil, D., Glowatz, H., & Schlumpberger, M. (2013). Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Current Protocols in Molecular Biology, 103*, 4.19.11-14.19.18.
Ohta, T. (2001). Neutral Theory. In S. Brenner & J. H. Miller (Eds.), *Encyclopedia of Genetics* (pp. 1326-1327). New York: Academic Press.
Orgel, L. E., & Crick, F. H. C. (1980). Selfish DNA - The ultimate parasite. *Nature, 284*, 604-607.
Palacios-Gimenez, O. M., Dias, G. B., de Lima, L. G., Kuhn, G. C. E. S., Ramos, E., et al. (2017). High-throughput analysis of the satellitome revealed enormous diversity of satellite DNAs in the neo-Y chromosome of the cricket *Eneoptera surinamensis*. *Scientific Reports, 7*, 6422.
Palestis, B. G., Trivers, R., Burt, A., & Jones, R. N. (2004). The distribution of B chromosomes across species. *Cytogenetic and Genome Research, 106*, 151-158.
Palomeque, T., & Lorite, P. (2008). Satellite DNA in insects: a review. *Heredity, 100*, 564-573.
Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*, 526-528.
Pasquesi, G. I. M., Adams, R. H., Card, D. C., Schield, D. R., Corbin, A. B., et al. (2018). Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nature Communications, 9*, 2774.
Peccoud, J., Loiseau, V., Cordaux, R., & Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences USA, 114*, 4721-4726.
Petersen, M., Armisen, D., Gibbs, R. A., Hering, L., Khila, A., et al. (2019). Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. *BMC Evolutionary Biology, 19*, 11.
Petitpierre, E. (1996). Molecular cytogenetics and taxonomy of insects, with particular reference to the coleoptera. *International Journal of Insect Morphology & Embryology, 25*, 115-134.
Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics, 17*, 23-28.
Petrov, D. A. (2002). Mutational equilibrium model of genome size evolution. *Theoretical Population Biology, 61*, 531-544.
Piednoel, M., Aberer, A. J., Schneeweiss, G. M., Macas, J., Novak, P., et al. (2012). Next-generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Molecular Biology and Evolution, 29*, 3601-3611.
Plohl, M., Luchetti, A., Meštrović, N., & Mantovani, B. (2008). Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene, 409*, 72-82.
Plohl, M., & Meštrović, N. (2012). Satellite DNA evolution. In M. A. Garrido-Ramos (Ed.), *Repetitive DNA* (pp. 126-152). Basel: Karger.
Plohl, M., Meštrović, N., & Mravinac, B. (2012). Satellite DNA Evolution. *Genome Dynamics, 7*, 126-152.

Qing, T., Yu, Y., Du, T., & Shi, L. (2013). mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. *Science China Life Sciences, 56*, 134-142.

Qiu, Z., Liu, F., Lu, H., & Huang, Y. (2016). Characterization and analysis of a *de novo* transcriptome from the pygmy grasshopper *Tetrix japonica*. *Molecular Ecology Resources*, 1-12.

Reinhold, K. (1999). Energetically costly behaviour and the evolution of resting metabolic rate in insects. *Functional Ecology, 13*, 217-224.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., et al. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature, 515*, 261-263.

Rowell, C. (1971). The variable coloration of the Acridoid grasshoppers. *Advances in Insect Physiology, 8*, 145-198.

Ruiz-Ruano, F. J., Cabrero, J., López-León, M. D., & Camacho, J. P. M. (2017). Satellite DNA content illuminates the ancestry of a supernumerary (B) chromosome. *Chromosoma, 126*, 487-500.

Ruiz-Ruano, F. J., Cabrero, J., López-León, M. D., Sánchez, A., & Camacho, J. P. M. (2018). Quantitative sequence characterization for repetitive DNA content in the supernumerary chromosome of the migratory locust. *Chromosoma, 127*, 45–57.

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports, 6*, 28333.

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual* (2nd edn. ed.). New York: Cold Spring Harbor Laboratory Press.

Schielzeth, H., Streitner, C., Lampe, U., Franzke, A., & Reinhold, K. (2014). Genome size variation affects song attractiveness in grasshoppers: Evidence for sexual selection against large genomes. *Evolution, 68*, 3629-3635.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics, 28*, 1086-1092.

Shah, A. B., Schielzeth, H., Albersmeier, A., Kalinowski, J., & Hoffman, J. I. (2016). High-throughput sequencing and graph-based cluster analysis facilitate microsatellite development from a highly complex genome. *Ecology and Evolution, 6*, 5718-5727.

Shapiro, J. A., & von Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biological Reviews, 80*, 227-250.

Smit, A. H., R. Green, P. (2015). RepeatMasker Open (Version 4.0). Retrieved from http://www.repeatmasker.org

Song, H., Amédégnato, C., Cigliano, M. M., Desutter-Grandcolas, L., Heads, S. W., et al. (2015). 300 million years of diversification: elucidating the patterns of orthopteran evolution based on comprehensive taxon and gene sampling. *Cladistics, 31*, 621-651.

Song, H., Mariño-Pérez, R., Woller, D. A., & Cigliano, M. M. (2018). Evolution, diversification, and biogeography of grasshoppers (Orthoptera: Acrididae). *Insect Systematics and Diversity, 2*, 193-125.

Song, H., Moulton, M. J., & Whiting, M. F. (2014). Rampant nuclear insertion of mtDNA across diverse lineages within Orthoptera (Insecta). *PLoS ONE, 9*, e110508.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*, 1312-1313.

Sword, G. A., & Simpson, S. J. (2000). Is there an intraspecific role for density-dependent colour change in the desert locust? *Animal Behaviour, 59*, 861-870.
Talla, V., Suh, A., Kalsoom, F., Dinca, V., Vila, R., et al. (2017). Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (*Leptidea*) butterflies. *Genome Biology and Evolution, 9*, 2491-2505.
Tanaka, S. (2004). Hormonal control of body-color polyphenism in the american grasshopper, *Schistocerca americana*: A Function of [His7]-Corazonin. *Journal of Insect Physiology, 46*, 1535-1544.
Thomas, J., & Pritham, E. J. (2015). Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiology Spectrum, 3*, MDNA3-0049-2014.
Uvarov, B. (1977). *Grasshoppers and Locusts. A Handbook of General acridology Vol. 2. Behaviour, ecology, biogeography, population dynamics*. London: Centre for Overseas Pest Research.
Valverde, J. P., & Schielzeth, H. (2015). What triggers colour change? Effects of background colour and temperature on the development of an alpine grasshopper. *BMC Evolutionary Biology, 15*, 168.
Vedenina, V., & Mugue, N. (2011). Speciation in gomphocerine grasshoppers: molecular phylogeny versus bioacoustics and courtship behavior. *Journal of Orthoptera Research, 20*, 109-125.
Vijay, N., Poelstra, J. W., Kunstner, A., & Wolf, J. B. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology, 22*, 620-634.
Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., et al. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications, 5*, 2957.
Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*, 57-63.
Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved from www.genome.gov/sequencingcostsdata
Wilhelm, B. T., & Landry, J.-R. (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods, 48*, 249-257.
Wilmore, P. J., & Brown, A. K. (1975). Molecular poperties of Orthopteran DNA. *Chromosoma, 51*, 337-345.
Wu, C., & Lu, J. (2019). Diversification of transposable elements in arthropods and its impact on genome evolution. *Genes, 10*, 338.
Wu, R., Wu, Z., Wang, X., Yang, P., Yu, D., et al. (2012). Metabolomic analysis reveals that carnitines are key regulatory metabolites in phase transition of the locusts. *Proceedings of the National Academy of Sciences USA, 109*, 3259.
Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., et al. (2014). SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics, 30*, 1660-1666.
Xiong, W. W., He, L. M., Lai, J. S., Dooner, H. K., & Du, C. G. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences USA, 111*, 10263-10268.
Yuan, Z., Liu, S., Zhou, T., Tian, C., Bao, L., et al. (2018). Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics, 19*, 141.

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

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*, 614-620.

Zytnicki, M., Akhunov, E., & Quesneville, H. (2014). Tedna: a transposable element *de novo* assembler. *Bioinformatics, 30*, 2656-2658.

Acknowledgements

A doctoral dissertation is the cumulative product of years of commitment, tears, blood, haemolymph, destroyed samples, crashed RAID arrays, well-aged RNA-later buffer solutions, many DNA and RNA extractions, successful and miserable field trips, “Error in data$V1 : $ operator is invalid for atomic vectors”, lab barbeques and perhaps a tinge of sheer mad passion to discover.

First and foremost, I would like to thank my doctoral advisers, Profs. Joe Hoffman and Holger Schielzeth for having the copious amounts of patience and empathy with me, and offering me this incredible opportunity. Their unyielding support was crucial for the completion of this dissertation. I must say, I have learnt considerably from both. Joe opened my eyes to the wonderful world of microsatellites, heterozygous fitness correlations and 90s gangster rap music (shout out to Coolio, DMX and Biggie). Holger introduced me to the incredible Acrididae grasshoppers, statistical approaches, and the nuanced appreciation of “vollkorn dopplekeks”. I am forever grateful to them for having confidence in my “crazy ideas” and “cunning stunts”.

My family has also been incredibly patient with me. First, I would like to thank my wife, Meha, for her unending support and understanding throughout my doctoral studies. Next, I would to thank my family, my parents, my brother Anirudh, my sister-in-law Toral, and my niece Anya, my nephew Abhimanyu, my grand mothers Subb-amma and Nani, and my in-laws, Ankush and Zam-Zam for supporting me throughout my career.

Next, I would like to recognize my wholesome and loving friends and colleagues at the Population Ecology group at the Institute of ecology and evolution, Ana, Alé, Anne, Anja, Anasuya, Günter, Gerlinde, Gabi, David C, Denise, Fabi, Max B, Max F, Raphael, Elina, Reto, Jutta, Ilka, Steffen, Ute, Katja and Silke for all their bratwürst, love, support, cakes and coffee during these years and for the future.

Next, I would like to recognize loving friends at the Animal Behaviour Department and Evolutionary Biology Department at Bielefeld University, Astrid, Athina, Anna, Anneke, Bahar, Pablo, Elke, David V, Emily, Katie, Luke, Lucy, Martin, Mathias, Meinolf, Micheal, Michele, Nayden, Stephie, Steffi, Stephie, Yumi, Tim, Steve, Peter, “Babsi”, Fritz, Ollie, and

Klaus. Thank You for all your love, support, guidance and intellectual banter during my time at the VHF!

I would also like to thank the IT support teams at the CeBiTec and iDiv compute clusters for entertaining my naïve questions. I would also like to the major funding agencies, DFG and the Marie Curie FP7-Reintegration Grant. Thank You for the generous support.

Last but not the least, a shout out to Rusty Hodge and the Big Earl for ‘keepin- spinning’ it at SomaFM for serving a nicely chilled plate of ambient downtempo beats and grooves with Groove Salad. Thank you for all the wonderful music.

This PhD would not have been possible without these wonderful people in my life!

THANK YOU ALL

Declaration of Independent Assignment

I declare in accordance with the conferral of the degree of doctor from the School of Biology and Pharmacy of the Friedrich-Schiller-University Jena that the submitted thesis was only written with the assistance and literature cited in the text.

People who assisted in experiments, data analysis and writing of the manuscripts are listed as co-authors of the respective manuscripts. I was not assisted by a consultant for doctorate theses.

This thesis has not been submitted whether to the Friedrich-Schiller-University, Jena or any other university.

Jena, August 10th, 2020 Abhijeet Shah