Leibniz Association

Final report

Title of the project:
Leibniz Graduate School on Genomic Biodiversity Research (GBR)

Leibniz-Institute: Zoological Research Museum Alexander Koenig
Reference number: SAW-2013-zfmk-4
Project period: 2013-2017
Contact partner: Prof. Dr. Bernhard Misof

Table of Contents

1. Background and aims

In the proposed Graduate School, we focused on the evolution of genes and genomes within holometabolous insects to understand the evolution of these insects in general and on the development and application of innovative bioinformatic tools that are needed to analyse genomic data. The primary research goal was to trace the evolution of holometabolous insect genomes and to infer a hypothetical ancestral genome of these animals with state of the art tools. In choosing holometabolous insects we could profit from the already published extensive genomic data and complement them with additional genome data for our specific aims.

Despite the availability of several sequenced genomes of holometabolous species, we are still unable to address fundamental questions of genome evolution within insects, such as: why are genomes of holometabolous insects usually smaller than those of hemimetabolous insects (insects with direct larval development)? In what other features do they differ from those of other insects? What genomic novelties evolved in the common ancestor of holometabolous insects that might have catalyzed their incredible success and diversification? How malleable is the order of genes in the genome? What is the role of domain rearrangements in the evolution of genes and gene function in insects? How did the ancestral genome of holometabolous insects look like? Do non-coding RNAs differ in their abundance and diversity within holometabolous insects? Is the enormous radiation of holometabolous insects correlated with gene duplications and subsequent functional diversification? Answers to these questions are currently lacking, but a considerate choice of holo- and hemimetabolous insects for de novo sequencing of their genomes in combination with published genomes would allow addressing the above questions. The intended set of fully sequenced genomes will furnish to deliver comparative descriptions of transposable elements, genome arrangements, evolution of novel genes and gene functions, and new potential markers for phylogenetics.

The inclusion of hemimetabolous insects is indispensable to identify evolutionary innovations of holometabolous insects at the genomic level. The genomes of the 10 species that we here propose to sequence combined with those that have already been sequenced (e.g., silk moth, flour beetle, mosquito, honey bee, fruit flies, jewel wasps, etc.) and newly sequenced genomes within the i5K consortium offered the potential to reach our scientific goals.

It was clear that the reliability of every evolutionary analyses depends on a robust reconstructions of phylogenetic relationships. With our contributions to 1KITE and i5k, we are in the fortunate situation to have exactly these at hand (international consortia on sequencing 1,000 insect transcriptomes, 1KITE (www.1kite.org), and 5,000 insect genomes, i5K (www.arthropodgenomes.org, Bernhard Misof as co-speaker of 1KITE, and Bernhard Misof and Oliver Niehuis as member of the i5K core unit). We could expect a robust phylogeny of holometabolous insects based on the extensive data produced within 1KITE. We were also be able to augment our genome data within the GBR with additional newly sequenced genomes and transcriptomes from these consortia.

The GBR also profited from our extensive cooperations in the field of genomics with Xin Zhou (Head of the National Bio-resource Bank at BGI, China) and Duane McKenna (Genomics at Memphis, University of Tennessee, USA).

With the advent of more efficient molecular technologies, systematics has also received a strong and extensive formalization of its theoretical foundations. In addition, a dramatic increase in computational power and genomic data has pushed molecular analyses into an area, in which theory and analysis is difficult, if not impossible, to comprehend for conventionally trained biologists. In all fields of research of systematics, extensive knowledge of (bio-)informatic and mathematical tools along with traditional training in biology is necessary to achieve an international competitive level of data analysis.

Table 1. Expertise of Principle Investigators:

Name	Institute	Organisms	Methods
Prof. Dr. B. Misof	ZFMK	Hexapoda (insects)	Phylogenomics, Molecular evolution
Prof. Dr. W. Wägele	ZFMK	Arthropoda	Phylogenomics, Phylogenetics (Theory)
Dr. O. Niehuis	ZFMK	Hexapoda	Comparative Genomics Phylogenomics
Dr. Ch. Mayer	ZFMK	--	Bioinformatics
Dr. R. Peters	ZFMK	Hexapoda	Phylogenomics
Dr. A. Donath	ZFMK	Hexapoda	Bioinformatics
Prof. Dr. Th. Bartolomaeus	IEZ	Arthropoda, Annelida (worms)	Phylogenomics, Phylogenetics (Theory)
PD Dr. L. Podsiadlowski	IEZ	Arthropoda	Phylogenomics, Comparative Genomics
Prof. Dr. J. Rust	Steinmann Inst.	Arthropoda	Phylogenetics, Palaeontology
Prof. Dr. E. Bornberg-Bauer	IEB	Hexapoda	Genome Analysis, Modular Protein Evolution
Dr. S. Grath	IEB	Hexapoda	Comparative Genomics

The cornerstone of the success of the GBR was a perfectly complementary expertise among the three principle groups executing the GBR (Table 1).

The cornerstone of the educational concept of the Graduate school was to develop a common language between biologists and bioinformaticians.

Students had access to modules of the OEP (Master of Science in Organismic Biology, Evolutionary Biology, and Palaeobiology) at the University of Bonn that covers basic to profound knowledge of the diversity (i.e. the morphology, anatomy, ecology, palaeontology, and evolution) of Metazoa. Additional courses cover the theory and methods of phylogenetic systematics, bioinformatics, and training in programming skills. In addition, students had access to modules of the Master program Special Study Program (SSP) in Evolution and Biocomplexity on comparative genomics, phylogenetics, and bioinformatics at the University of Münster.

The total duration of the GBR wias four years, ending in April 2017. However, each PhD student was going to be funded for only three years within the GBR. We tried to extend the duration of the PhD project by in-house or follow-up DFG funding up to four/five years. Part of the results described below were finally achieved after the original funding period of the PhD students.

2. Execution of the project, deviations from original aims, technical problems

We were able to appoint all applied for student positions until October 2014, except for three: (1) one PhD student position had to be canceled, because of funding cuts, (2) one PhD student position was funded by inhouse money from the ZFMK, in order not loose necessary research aspects, and (3) one PhD student was appointed after October 2014.

The principle investigators of the GBR decided to complement published genome sequence assemblies with de novo characterized genome sequence assemblies of 30 additional species, mostly hymenopteran. This species selection gave us the opportunity to study genome evolution within insects in great detail (Table 2).

The different PhD projects were concerned with different aspects of genome evolution and characteristics (Table 3):

(1) The evolution of the gene repertoire

(2) The evolution of the mobilome

(3) The evolution of the non-coding repertoire

(4) The evolution of extreme genome size and Hymenoptera genomes

(5) The origin of extremely small protein (SMORFs)

(6) The evolution of the domain content

(7) Calibration the evolution of genomes

Table 3 PhD Students and Projects

Name	Institute	Project	PI
Tanja Ziesmann	ZFMK	Project 3	Donath, Misof
Jeanne Wilbrandt	ZFMK	Project 1	Niehuis, Misof
Malte Petersen	ZFMK	Project 2	Misof, Mayer
Jan-Philip Oeyen	ZFMK	Project 4	Misof, NIehuis
Simon Gunkel	Steinmann Inst.	Project 7	Rust, Wappler
Steffen Klassberg	U. Münster	Project 6	Bornberg
Peter Lesny	U. Bonn	Project 5	Podsiadlowski

All of these projects were based on published and de novo characterized genome assemblies. We finally skipped the project of reconstructing the ancestral holometabolous genome characteristics, because of time constraints and the restrictive quality of the genome assemblies.

It turned out that most of the published genome sequence assemblies were of restricted quality in terms of scaffold length limiting the analyses of the mobilome, gene repertoire and non-coding elements, however still perfectly fine to analyse domain structure. Since the Pis of the project are members of the i5k consortium, we had prepublication access to other

unpublished holometabolous genome sequence assemblies which extended our taxon sampling, but still did not solve the quality restrictions.

Our de novo sequencing was successful in delivering over 20 well characterized genome sequence assemblies, based on Illumina sequencing in cooperation with the BGI (Table 2). However, these genome assemblies suffered quality restrictions as well, since based exclusively on Illumina techniques. Today, the enormous advances with the PacBio and Oxford Nanopore techniques offer a chance to dramatically improve genome sequence assemblies.

For a selection of GBR species we have now complemented the Illumina based assemblies with long-read data from Oxford Nanopore, using inhouse money (ZFMK). The combined assemblies improved dramatically and we are now in a situation to further analyse these genomes. The PhD projects were thus executed on a limited set of genome sequence assemblies, but still with an interesting potential for general results.

3. Results and scientific relevance

Despite the mentioned technical problems, we were able to proceed with the analyses of gene repertoires, domains, mobilomes, non-coding elements and the analyses of genome size evolution.

3.1. The evolution of gene repertoires

The evolution of gene repertoires and their impact on genome size evolution was the cornerstone of this project. Before this project, it was unclear whether or not gene repertoire dynamics leaves any footprint in the evolution of genome size. Additionally, it was unclear whether or not automatic genome annotation delivers data of sufficient quality to study these phenomena. Manual curation is still considered the gold standard of gene model annotation, therefore it was important to asses the effect of manual annotation on structural gene parameters and gene annotation quality.

We compared five selected structural properties ([i] unspliced transcript length, [ii] protein length, [iii] exon count per transcript, as well as [iv] median exon and [v] median intron length per transcript) of protein-coding genes of automatically annotated gene sets with those of manually annotated gene sets of seven non-model insect species sequenced by the i5k initiative (i5k Consortium 2014). Gene structural properties were assessed with the tool COGNATE (Wilbrandt et al. 2017).

The results show that the properties of automatically generated and manually curated gene models differ only marginally from subsets that were manually annotated (Fig. 1A). Furthermore, major trends regarding gene structure properties can be inferred from both automatically predicted gene sets and manually annotated gene models alike (Wilbrandt et al., under revision in BMC Genomics).

We partitioned gene repertoires into classes according to the gene's orthology relationships and conservation. We called these classes "core", "shell", and "cloud". The core partition includes all gene families present at node 1 (see Fig. 1B for phylogeny and node IDs). The shell consists of all gene families not present in the core but at the following nodes that comprise more than two species (nodes 2–8, 12–16, 18–20, 22–24). The cloud comprises all gene families present in exactly two sister species (five nodes: 9, 10, 17, 21, 25) or only in one species (nonOG). Based on the count of gene family members four states are discriminated:

- USCs (universal single-copy orthologs): all species descending from the considered node have exactly one copy of the considered gene family.

- sSCs (species-specific single-copy orthologs): the considered species contributes exactly one copy.
- sMCs (species-specific multicopy): the considered species contributes more than one copy.
- NonOG (not a member of any ortholog group): no orthologous gene is found in any other species of the sample. (Not necessarily young genes in cases where one species represents an old clade, e.g., only Sirex noctilio represents Siricoidea, which split 264 mya from the next relatives in the analyzed phylogeny [Misof et al. 2014].)

A B

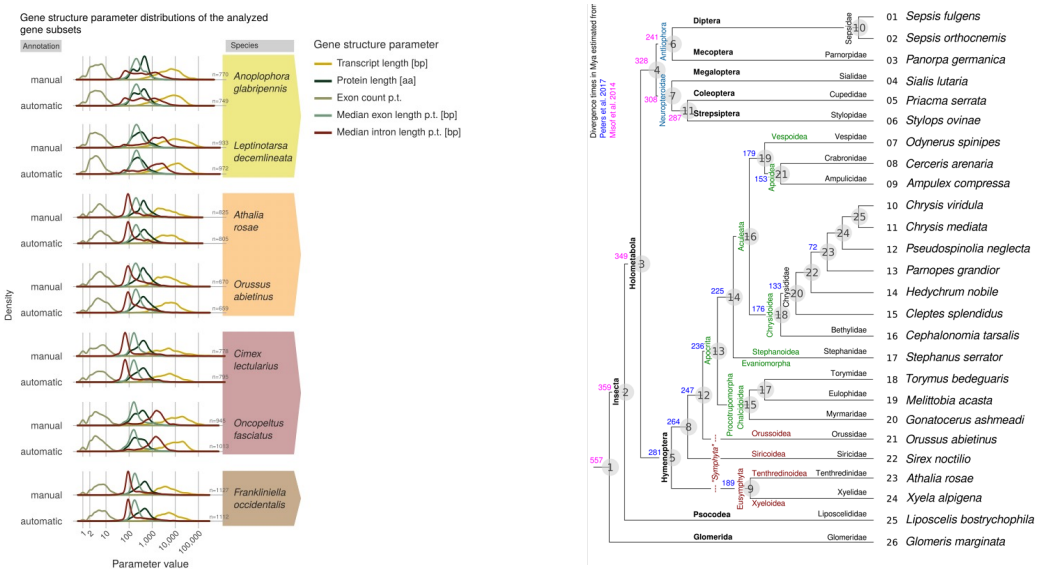


Figure 2: (A) Density distributions of five gene structure properties per genome (semi-logarithmic): unspliced transcript length [bp], protein length [aa], exon count p.t., median exon length p.t. [bp], median intron length p.t. [bp]. Ridge panels show the two sets (top row: manually curated subset, bottom row: corresponding automatically generated subset) for each of the seven species (Anoplophora glabripennis [Coleoptera], Athalia rosae [Hymenoptera], Cimex lectularius [Hemiptera], Frankliniella occidentalis [Thysanoptera], Leptinotarsa decemlineata [Coleoptera], Oncopeltus fasciatus [Hemiptera], Orussus abietinus [Hymenoptera]). aa: amino acids; bp: base pairs; Mbp: mega base pairs; p.t.: per transcript. (B) Phylogenetic relationships between species of the sample. Numbers in grey circles are node IDs. Each species has a unique ID prepending its name. Pink numbers indicate estimated approximate divergence times in Mya (Million years ago) referring to Misof et al. (2014), while blue numbers mark estimated approximate divergence times from the publication by Peters et al. (2017). Coloring of taxonomic names follows the scheme red for “Symphyta” and the comprised lineages, green for Apocrita and descending lineages, and blue for other Holometabola than Hymenoptera. Branch lengths are arbitrary. Mya: million years ago.

Species differ tremendously in their gene counts; the contribution to gene count by genes of the three conservation classes and four copy states is also variable (Fig. 3, left side). All analyzed hymenopteran gene repertoires contain over 90% of the complete benchmark single-copy orthologs (BUSCOs, Simão et al. 2015). Aculeata have similar repertoire sizes of less than 20,000 genes. However, genome size does not correlate with gene count (but with

TE content; see below). The predicted gene models cover the expected gene space very well, especially within studied Hymenoptera.

Core, Shell, and Cloud conservation classes differ in gene structure and domain diversity

Genes of the core class have longer transcripts and proteins, but shorter exons and introns. Universal single-copy orthologs (core: USC) represent the longest genes and have most introns. We find that genes classified as core are generally longer and produce longer proteins due to a higher complexity (more, but relatively short exons and introns) than lineage-specific genes (cloud). This corroborates the 'universal length difference' hypothesis (e.g., Clark et al. 2007; Yang et al. 2013; but also see Tatarinova et al. 2016). The correlation of protein length and conservation has been accredited to functional relevance (Lipman et al. 2002), but the origin and maintenance of the observed patterns are unclear.

It can be expected that aging genes (i.e., genes that are retained after their origination) become longer over time. It rests on the assumption that new genes are less likely to be long and/or complex when originating (Wissler et al. 2013). However, it can also be assumed that there are natural limits to gene growth, imposed by physical limits and by cost limitations related to transcription, replication, or product toxicity (e.g., Drummond and Wilke 2008). It has been suggested that intron gain is adaptive (Carmel et al. 2007), thus playing a role in the elongation of retained genes.

One hypothesis is that core genes are (functionally) important, otherwise they would not be conserved over very large time scales (e.g., Jordan et al. 2002). The high complexity and length of these apparently important genes has benefits and drawbacks. One advantage is the possibility to produce more alternative splice variants, improving the ratio of proteins to required nucleotide sequence length (Roux and Robinson-Rechavi 2011 , Kianianmomeni et al. 2014). A disadvantage of high complexity and length is the theoretically higher probability of deleterious mutations and insertions of TEs. Furthermore, the sheer length of a gene theoretically also impacts transcription speed and accuracy (Castillo-Davis et al. 2002). Additionally, it has been found that the most conserved orthologs show also the highest DNA methylation levels in insects, potentially playing a role in reducing transcriptional noise (Provataris et al. 2018). This is in line with the suggestion that DNA methylation and gene expression regulation are interconnected and (partially) drive gene length evolution (Zeng and Yi 2010).

In total 27 % of all genes were annotated with one or more protein domains. Of these, 74.8 % were assigned to the core class, 18.2 % to shell genes, and 6.9 % to cloud genes (94.7 % of these are nonOG genes). The ratio of domains per gene is highest in the core: 1.83 domains/gene (shell: 1.52; cloud: 1.28). We assessed the diversity of protein domains and three kinds of domain arrangements (pairs, triplets, quartets) between the three conservation classes. The ratio of unique domains per gene is highest in the cloud (0.54; shell: 0.13; core: 0.05). Regarding individual protein domains, a high diversity can be observed in all three conservation classes, although much of the shell's diversity is shared, while most of the diversity found in core and cloud is specific to these classes. Domains found in the core can confidently considered to be old, their origin most likely dates back to at least 557 million years ago (Fig. 1B). Almost two thirds of these old domains are also found in the cloud. Our observations contradict the previously reported pattern (Gabaldón 2005) of most protein domains being ancient, while most combinations being lineage-specific.

3.2. The evolution of the mobilome

We annotated transposable elements (TEs) in 73 arthropod genomes (thus exploiting a much larger sample than initially planned) by using a combination of reference-based and de novo TE prediction approaches. The annotation procedure also included a rigorous filtering step to avoid overestimating the TE content. We found that genome size is linearly correlated with TE abundance. Furthermore, genome size is also dependent on the diversity of the TE

repertoire. These findings suggest that TE activity not only plays a large role in genome size evolution, but also that the more different TE superfamilies are present in a genome, genomic defenses against TE proliferation are less efficient.

To visualize the TE content in more detail, we generated a so-called repeat landscape for each species (Fig. 4). The visualizations suggest in many cases that TEs invaded the insect genomes in bursts (high rate of TE proliferation in a short period of time). Using the TE annotations from our pipeline, we mapped each TE copy's occurrence in a dated phylogenetic tree and classified the TE copies into lineage-specific and ancestral TEs. The data reveal that in a clade that is more than about 120 million years old, no ancestral TEs could be identified. This is likely a result of both targeted silencing/inactivation and random mutations that degrade the DNA sequence of TEs over time.

The generated data from this project are currently being used in one of the PIs group's (B.M.) in a study on DNA methylation patterns in TEs within genes. Furthermore, the data provide a solid starting point for detailed studies on TE evolution in many insect orders and have been downloaded multiple times by other researchers since they became available: datadryad.org/resource/doi:10.5061/dryad.55p667b/1

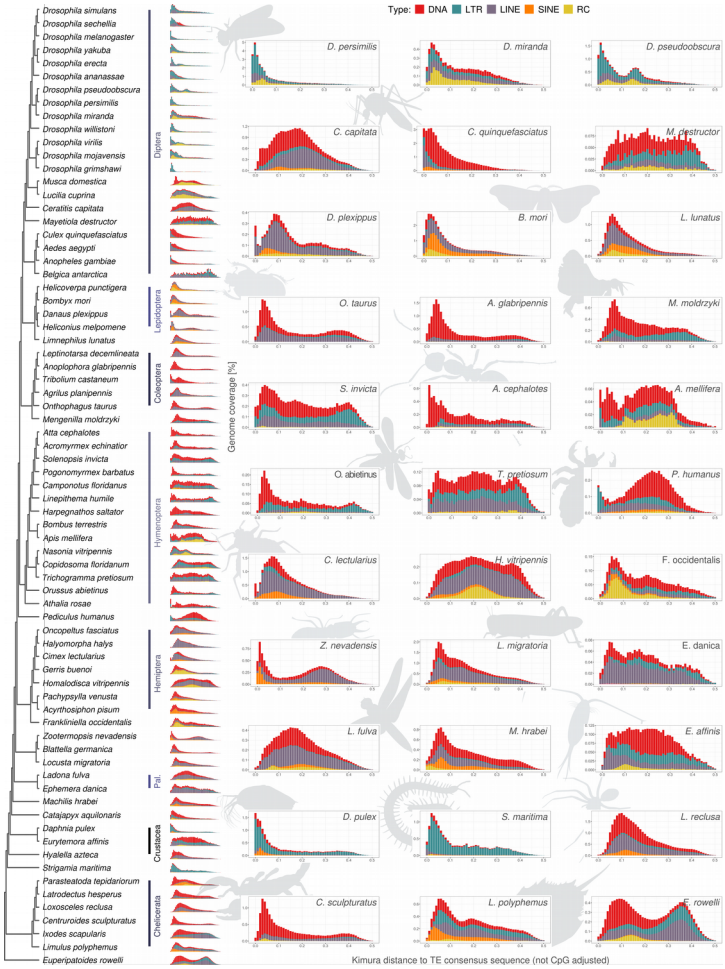


Figure 3: Cladogram with repeat landscape plots. The larger plots are selected representatives. The further to the left a peak in the distribution is, the younger the corresponding TE fraction generally is (low TE intra-family sequence divergence). In most orders, the TE divergence distribution is similar, such as in Diptera or Hymenoptera. The large fraction of unclassified elements was omitted for these plots. Pal., Palaeoptera. Figure from Petersen et al. (2019).

3.3. The evolution of the non-coding repertoire

The annotation of the non-coding repertoire posed severe problems, as expected. Therefore, in order to generate meaningful data we restricted our analyses on the annotation of the non-coding repertoire of Hymenoptera. It turned out that many of classically described small RNA family genes can be identified in Hymenoptera and deliver important insights into the evolution of this group. In particular the long non-coding RNA genes turned out to be realing interesting, as they appear to be associated with so-called clusters of conserved non-coding elements (CNE). CNEs have not been systematically studied in insects and only little is speculated concerning their function in vertebrates. In vertebrates, CNEs seem to be associated with transcription factors and might be important regulators of their expression. In contrast, we found a clear non-random association between clusters of CNEs and long non-coding RNAs, which are in part conserved among hymenopterans. It can be speculated that this relationship between CNEs and long non-coding RNAs is related to gene regulation, as it was alread hypothesized that long non-coding RNAs might be involved in gene regulation as well.

Currently two publications are in preparation on these findings. They are also part of a large consortial analyses of the evolution of hymenopteran genomes lead by another PhD student of the GBR.

3.4. The evolution of domain content

The evolution of domain content, which can be losely synonimized with the evolution of functionality in genomes was analysed on the selected set of hymenopteran genomes as well. The results of this analyses are part of the hymenopteran genome analyses and being prepared for publication herein.

3.5. The evolution of extreme genome size and the evolution of hymenopteran genomes

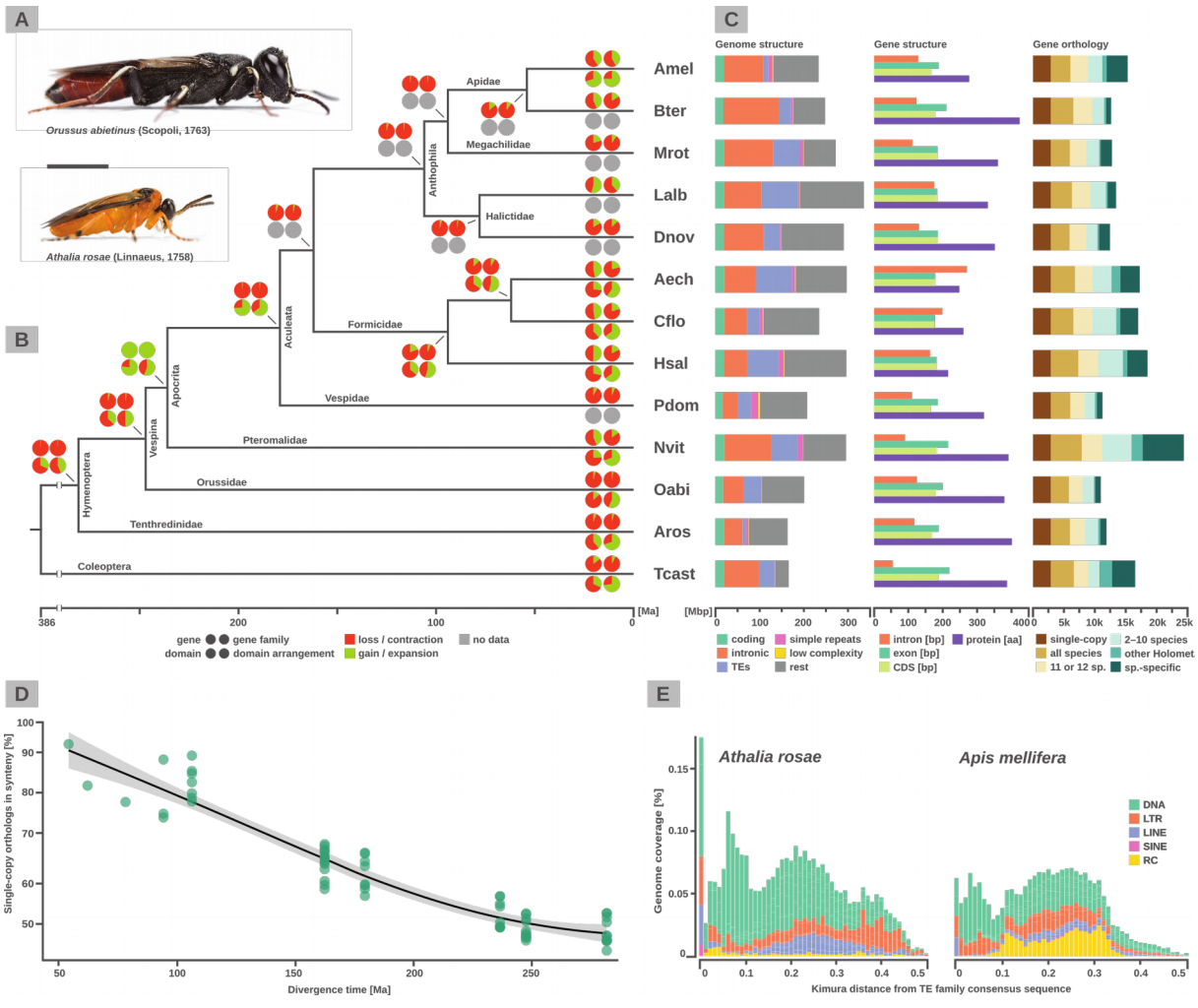
Within the GBR we sequenced the currently smallest known insect genome of about 50 Mb. This genome size estimate is based on kmer-estimates, and total genome assembly size of Illumina and Oxford Nanopore data. The PhD student (starting after October 2014) is focusing on the analyses of this extremely small genome in addition to the analyses of the evolution of the hymenopteran genomes. He is now funded from ZFMK inhouse money until 2020. With the late addition of Oxford Nanopore reads, we were able to assemble a high quality genome of Stylops ovinae which indicates that this genome has lost parts of its gene repertoire and seems to have reduced its size by extreme reduction of non-coding elements and sequences.

Additionally, the comparative analyses of hymenopteran genome sequence assembly delivered insights into the evolution of hymenopteran life history traits. Within Hymenoptera, we have plesiomorphic phytophagous groups the so-called „Symphyta“ and derived parasitoids, constituting the extremely species-rich clade of hymenopterans. The Orussidae, a species-poor clade, shows anatomical features of Symphyta but life history traits of parasitoids. We therefore analysed whether or not the Orussidae do show a genomic mosaic as well. It turned out that is exactly the case and Orussidae clearly show adaptations to the parasitoid life style besides plesiomorphic features in their genomes (Figure 4).

3.6. Calibrating the evolution of genomes

The necessary prerequiste of the evolutionary interpretations, namely a robust backbone tree of holometabolous insects was generated within the 1KITE consortium with contributions from PIs and PhD students of this project (Misof et al. 2014). In this phylogenetic project, the PhD students developed their phylogenomic expertise, contributed software and empirical analyses at an internationally competitive level. Additional to this project we developed a new approach for calibrating evolutionary events as a prerequiste for proper calibration of genome evolution.

The approach rests on the idea that different calibration regimes can be evaluated accoeding to their fit to the fossil record. A maximum likelihood criterion was developed for this approach and the entire procedure realized in a software package (Gunkel et al., 2017).



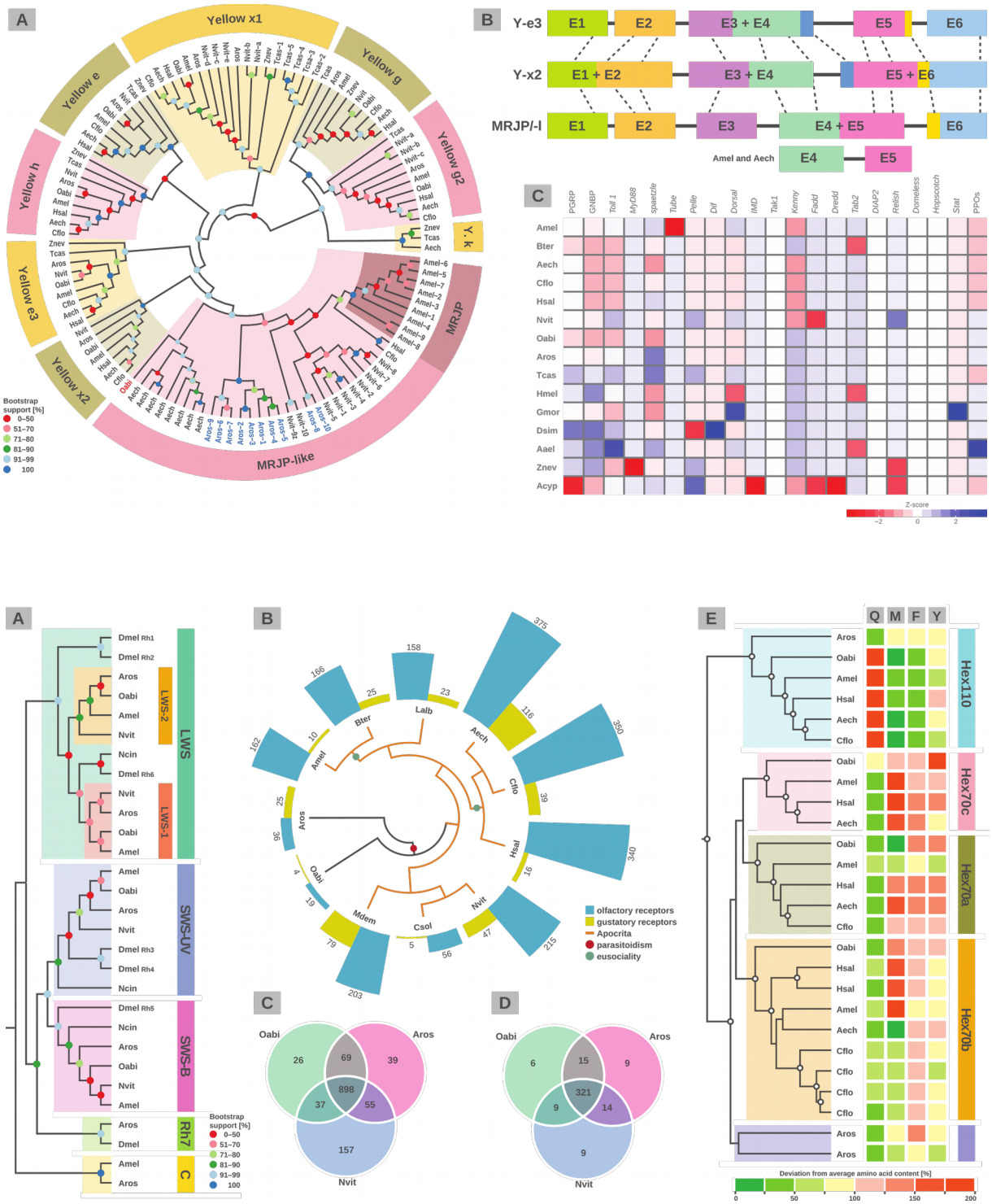


Figure 4. Structural features of compared hymenopteran genomes (from Oeyen et al., in prep).

3.7. References

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. (2015): The Octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 524: 220–224.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006): The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7: R43.

Carmel L, Rogozin IB, Wolf YI, Koonin EV (2007): Evolutionarily conserved genes preferentially accumulate introns. Genome Res. 17: 1045–1050.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nat Genet. 31: 415–418.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. (2007): Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.

Consortium i5K (2003): The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J. Hered. 104: 595–600.

Drummond DA, Wilke CO (2008): Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Gabaldón T (2005): Evolution of proteins and proteomes: a phylogenetics approach. Evol. Bioinform. Online 1:117693430500100000.

Hahn MW, Han MV, Han S-G (2007): Gene family evolution across 12 Drosophila genomes. PLoS Genet 3: e197.

Hahn MW, Bie TD, Stajich JE, Nguyen C, Cristianini N (2005): Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15: 1153–1160.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002): Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12: 962–968.

Kianianmomeni A, Ong CS, Rätsch G, Hallmann A (2014): Genome-wide analysis of alternative splicing in Volvox carteri. BMC Genomics 15: 1117.

Li Z, Tiley G, Galuska S, Reardon C, Kidder T, Rundell R, et al. (2018): Multiple large-scale gene and genome duplications during the evolution of hexapods. BioRxiv: 253609.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA (2002): The relationship of protein conservation and sequence length. BMC Evol. Biol. 2: 20.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346: 763–7.

Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, et al. (2017): Evolutionary history of the Hymenoptera. Current Biology 27: 1013–1018.

Provataris P, Meusemann K, Niehuis O, Grath S, Misof B, Wagner G (2018): Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. Genome Biol. Evol. 10: 1185–1197.

Rödelsperger C, Streit A, Sommer RJ (2013): Structure, function and evolution of the nematode genome. Chichester, UK: John Wiley & Sons, Ltd.

Roux J, Robinson-Rechavi M (2011): Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. Genome Res. 21: 357–363.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015): BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Tatarinova TV, Lysnyansky I, Nikolsky YV, Bolshoy A (2016): The mysterious orphans of Mycoplasmataceae. Biology Direct 11: 2.

Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, …, Misof B, …, Niehuis O, et al. (2018): The Genomic Basis of Arthropod Diversity. BioRxiv 382945.

Wang Z, Zarlenga D, Martin J, Abubucker S, Mitreva M (2012): Exploring metazoan evolution through dynamic and holistic changes in protein families and domains. BMC Evol. Biol. 12: 138.

Waterhouse RM (2015): A maturing understanding of the composition of the insect gene repertoire. Current Opinion in Insect Science 7: 15–23.

Wilbrandt J, Misof B, Niehuis O (2017): COGNATE: comparative gene annotation characterizer. BMC Genomics 18: 535.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E (2013): Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol. Evol. 5: 439–455.

Yang L, Zou M, Fu B, He S (2013): Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. BMC Genomics 14: 65.

Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, et al. (2008): Repetitive element-mediated recombination as a mechanism for new gene origination in Drosophila. PLoS Genet 4: e3.

Zeng J, Yi SV (2010) DNA methylation and genome evolution in honeybee: gene length, expression, functional enrichment covary with the evolutionary signature of DNA methylation. Genome Biol. Evol. 2: 770–780.

4. Collaborations

Collaborators within Germany

Georg Mayer, Department of Zoology, Institute of Biology, University of Kassel, Heinrich-Plett-Str. 40, 34132, Kassel, Germany

Lars Hering, Department of Zoology, Institute of Biology, University of Kassel, Heinrich-Plett-Str. 40, 34132, Kassel, Germany

Collaborators in foreign countries

Dr. David Armisén, Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, 69364, Lyon, France

Prof. Dr. Richard Gibbs, Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, 77030, Houston, TX, USA

Dr. Abderrahman Khila, Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, 69364, Lyon, France

Dr. Kristen A. Panfilio, School of Life Sciences, University of Warwick, Gibbet Hill Campus, Coventry CV4, 7AL, United Kingdom

Prof. Dr. Stephen Richards, Human Genome Sequencing Center, Department of Human and Molecular Genetics, Baylor College of Medicine, 77030, Houston, TX, USA

Prof. Dr. Robert M. Waterhouse, 3211 Le Biophore, Ecology & Evolution, Université de Lausanne, 1015 Lausanne, Switzerland

Members of the i5k consortium

Members of the 1KITE consortium

5. Publications

published manuscripts

Wilbrandt J, Misof B, Niehuis O (2017): COGNATE: comparative gene annotation characterizer. BMC Genomics 18: 535.

Petersen M, Armisen D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B (2019): Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Evolutionary Biology 19: 1

Gunkel S, Rust J, Wappler T, Mayer Ch, Niehuis O, Misof B (2017). A Posteriori Evaluation Of Molecular Divergence Dates Using Empirical Estimates Of Time-Heterogeneous Fossilization Rates. bioRxiv : 128314;doi: https: //doi.org/10.1101/128314.

submitted manuscripts

Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M, Binford GJ, Chao H, Chen YH, Childers C, Dinh H, Doddapaneni HV, Duan JJ, Dugan S, Esposito LA, Friedrich M, Garb J, Gasser RB, Goodisman MAD, Gundersen-Rindal DE, Han Y, Handler AM, Hatakeyama M, Hering L, Hunter WB, Ioannidis P, Jayaseelan JC, Kalra D, Khila A, Korhonen PK, Lee CE, Lee SL, Li Y, Lindsey ARI, Mayer G, McGregor AP, McKenna DD, Misof B, Munidasa M, Munoz-Torres M, Muzny DM, Niehuis O, Osuji-Lacy N, Palli SR, Panfilio KA, Pechmann M, Perry T, Peters RS, Poynton HC, Prpic N-M, Jiaxin Qu, Rotenberg D, Schal C, Schoville SD, Scully ED, Skinner E, Sloan DB, Stouthamer R, Strand MR, Szucsich NU, Wijeratne A, Young ND, Zattara EE, Benoit JB, Zdobnov EM, Pfrender ME, Hackett KJ, Werren JH, Worley KC, Gibbs RA, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S (in review after resubmission): The genomic basis of arthropod diversity. Nature [manuscript attached as PDF; initially submitted version also available at https://www.biorxiv.org/content/10.1101/382945v1)

Wilbrandt J, Misof B, Panfilio K, Niehuis O (in revision): Repertoire-wide gene structure analyses: a case study comparing automatically predicted and manually annotated gene models – BMC Genomics

manuscripts ready for submission

Oeyen JP, Benoit JB, Beukeboom LW, Bornberg-Bauer E, Buttstedt A, Cash EI, Chao H, Chen M-J, Childers C, Cridge AG, Dearden P, Dinh H, Doddapaneni HV, Dolan A, Donath A, Dowling D, Dugan S, Duncan E, Elpidina EN, Friedrich M, Geuverink E, Gibson JD, Grath S, Grimmelikhuijzen CJP, Große-Wilde E, Gudobba C, Han Y, Hansson BS, Hauser F, Hughes DST, Ioannidis P, Jacquin-Joly E, Jennings EC, Jones JK, Klasberg S, Lee SL, Lesný P, Lovegrove M, Martin S, Martynov AG, Mayer C, Montagné N, Moris VC, Munoz-Torres M, Murali SC, Muzny DM, Oppert B, Pauli T, Peters RS, Petersen M, Pick C, Podsiadlowski L, Poelchau MF, Provataris P, Qu J, von Reumont BM, Rosendale AJ, Simao FA, Skelly J, Sotiropoulos AG, Stahl AL, Sumitani M, Szuter EM, Tidswell O, Tsitlakidis E, Vedder L, Waterhouse RM, Werren JH, Wilbrandt J, Worley KC, Yamamoto DS, van de Zande L, Zdobnov E, Ziesmann T, Gibbs RA, Richards S, Hatakeyama M, Misof B, Niehuis O (in prep.): Draft genomes of two sawflies reveal evolutionary acquisitions that fostered the mega-radiation of parasitoid and eusocial Hymenoptera. [manuscript attached as PDF. Manuscript was in the attached version submitted to Current Biology and was rejected; we are currently revising the manuscript and intend to submit it to Nature Communications]

Petersen M, Nottebrock G, Misof B (in prep.): Dynamics of genome size evolution in arthropods with particular focus on insects. [early stage manuscript draft]

Oral and poster presentations resulting from the project

Lesny P, Petersen M, Meusemann K, Zhou X, Donath A, Niehuis O, Vilcinskas A, Misof B, Zhou X, Podsiadlowski L (2013). Antimicrobial peptide diversity among Hexapoda as revealed by a detailed survey of transcriptomic data from more than 300 species. DZG Tagung München 2013 (poster)

Martin S, Meusemann K, Petersen M, Misof B, Zhou X, Vilcinskas A, Podsiadlowski L (2013). Comparative transcriptomic approaches reveal a complex evolutionary history of matrix metalloproteinases in Diptera. DZG Tagung München, 2013 (talk)

Niehuis O (2016): Elucidating the evolutionary history and ecology of insects. Highlights from research on ants, bees and (parasitic) wasps as well as exploitation of NGS technologies. — University of Greifswald, Greifswald (June 21) [invited talk]

Niehuis O (2017): Die Evolutionsgeschichte der Hautflügler: eine Zeitreise mithilfe der vergleichenden Genomik. — Dies Academicus der Albert-Ludwigs-Universität Freiburg (July 28) [inaugural lecture]

Wilbrandt J (2016): Current and future projects. GBR and DFG grant. Extent, speed, and causes of changes in the protein-coding gene repertoire of holometabolous insects. — Leibniz Graduate School on Genomic Biodiversity Research round table, Zoological Research Museum Alexander Koenig, Bonn (July 6). [talk]

Wilbrandt J (2016): Annotation Characterization and Plot Preparation. ACPP: A tool for comparative genomics. — Leibniz Graduate School on Genomic Biodiversity Research round table, Zoological Research Museum Alexander Koenig, Bonn (September 9). [talk]

Wilbrandt J (2017): The protein-coding gene repertoire of insects. A PhD in comparative genomics. — Retreat of the PhD students at the Zoological Research Museum Alexander Koenig, Bonn (March 10). [talk]

Wilbrandt J (2017): Gene repertoires in insects. Approaches and insights. — Organismic Zoology Conference, University of Bonn, Bonn (May 20). [talk]

Wilbrandt J, Misof B, Niehuis O (2017): Characterizing gene repertoires or discover your music. — N2 Science Communication Conference at the Museum für Naturkunde Berlin, Berlin (November 6–8). [poster]

Wilbrandt J, Misof B, Panfilio K, Niehuis O (2017): Data basis, tool choice, human review: influences on predicted protein-coding gene structure. — 110th Annual Conference of the German Zoological Society, University of Bielefeld, Bielefeld (September 12–15). [poster]

Petersen M (2016): Transposable elements and tandem repeats shape insect genome evolution. — Leibniz Graduate School on Genomic Biodiversity Research round table, Zoological Research Museum Alexander Koenig, Bonn (June 2016). [talk]

Petersen M (2017): Transposable elements and tandem repeats shape insect genome evolution. — Retreat of the PhD students at the Zoological Research Museum Alexander Koenig, Bonn (March 10). [talk]

Petersen M (2018): Automated annotation of transposable elements. — Leibniz Graduate School on Genomic Biodiversity Research round table, Zoological Research Museum Alexander Koenig, Bonn (May 2018). [talk]

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, and Misof B (2018): Diversity and Evolution of the Transposable Element Repertoire in Insects. — Joint Congress on Evolutionary Biology, Montpellier (August 2018). [poster]

Petersen M, Oeyen JP, Wilbrandt J, Ziesmann T (2016): Investigating Biodiversity on a Genomic Scale. — Leibniz Evaluation at the Zoological Research Museum Alexander Koenig (September 2016). [poster]

6. Public Access to Genome Data

The characterized genome sequence assemblies are available upoin request from the GBR contact partner and will be fully available to the public upon publications of the manuscripts as it is mandatory for genome papers.