Published online 30 September 2019

doi: 10.1093/nargab/lqz006

A comprehensive annotation and differential expression analysis of short and long non-coding RNAs in 16 bat genomes

Nelly F. Mostajo[1,2,3,†], Marie Lataretu[1,3,†], Sebastian Krautwurst[1,3], Florian Mock[1,3], Daniel Desirò[1,3], Kevin Lamkiewicz[1,3], Maximilian Collatz[1,3], Andreas Schoen[4,5], Friedemann Weber[2,4,5], Manja Marz[1,3,6,†] and Martin Hölzer[1,3,*,†]

[1]RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany, [2]Institute of Virology, Philipps-University Marburg, Hans-Meerwein-Straße 2, 35043 Marburg, Germany, [3]European Virus Bioinformatics Center, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany, [4]Institute for Virology, FB10-Veterinary Medicine, Justus-Liebig University, 35392 Gießen, Germany, [5]German Center for Infection Research (DZIF), partner sites 35043 Marburg and 35392 Gießen, Germany and [6]FLI Leibniz Institute for Age Research, Beutenbergstraße 11, 07745 Jena, Germany

Received July 03, 2019; Revised August 21, 2019; Editorial Decision September 05, 2019; Accepted September 10, 2019

ABSTRACT

Although bats are increasingly becoming the focus of scientific studies due to their unique properties, these exceptional animals are still among the least studied mammals. Assembly quality and completeness of bat genomes vary a lot and especially non-coding RNA (ncRNA) annotations are incomplete or simply missing. Accordingly, standard bioinformatics pipelines for gene expression analysis often ignore ncRNAs such as microRNAs or long antisense RNAs. The main cause of this problem is the use of incomplete genome annotations. We present a complete screening for ncRNAs within 16 bat genomes. NcRNAs affect a remarkable variety of vital biological functions, including gene expression regulation, RNA processing, RNA interference and, as recently described, regulatory processes in viral infections. Within all investigated bat assemblies, we annotated 667 ncRNA families including 162 snoRNAs and 193 miRNAs as well as rRNAs, tRNAs, several snRNAs and lncRNAs, and other structural ncRNA elements. We validated our ncRNA candidates by six RNA-Seq data sets and show significant expression patterns that have never been described before in a bat species on such a large scale. Our annotations will be usable as a resource (rna.uni-jena.de/supplements/bats) for deeper studying of bat evolution, ncRNAs repertoire, gene expression and regulation, ecology and important host–virus interactions.

INTRODUCTION

Bats (Chiroptera) are the most abundant, ecologically diverse and globally distributed animals within all vertebrates (1), but representative genome arrangements and corresponding coding and non-coding gene annotations are still incomplete (2). Except for the extreme polar regions, bats can be found throughout the globe, feeding on diverse sources such as insects, blood, nectar, fruits and pollen (2). Their origin has been dated in the Cretaceous period, with a diversification explosion process dating back to the Eocene (3).

The 21 bat families known to date are classified into the suborders *Yinpterochiroptera* and *Yangochiroptera* (2,4) (Figure 1). Although they account for >20% of the total living mammalian diversity (5), the genomes of only 16 bat species of the estimated >1300 species (2) have been sequenced with adequate coverage to date and are publicly available (Figure 1).

Bats have developed a variety of unique biological features that are the rarest among all mammalian, including laryngeal echolocation (4,6), vocal learning (7) and the ability to fly (3). They occupy a broad range of different ecological niches (2), have an exceptional longevity (8–10) and a natural and unique resilience against various pathogenic viruses (1,11). For example, bats are the suspected reservoirs for some of the deadliest viral diseases such as Ebola and SARS (12–14), but appear to be asymptomatic and survive the infection. Possibly, the solution to better understand and fight these pathogens lies in the uniquely developed immune system of bats (15,16). Studying bats and their genomes is likely to have high impacts on various sci-

*To whom correspondence should be addressed. Tel: +49 3641 9 46485; Email: martin.hoelzer@uni-jena.de

†The authors wish it to be known that, in their opinion, the first and the last two authors should be regarded as joint first/last authors.

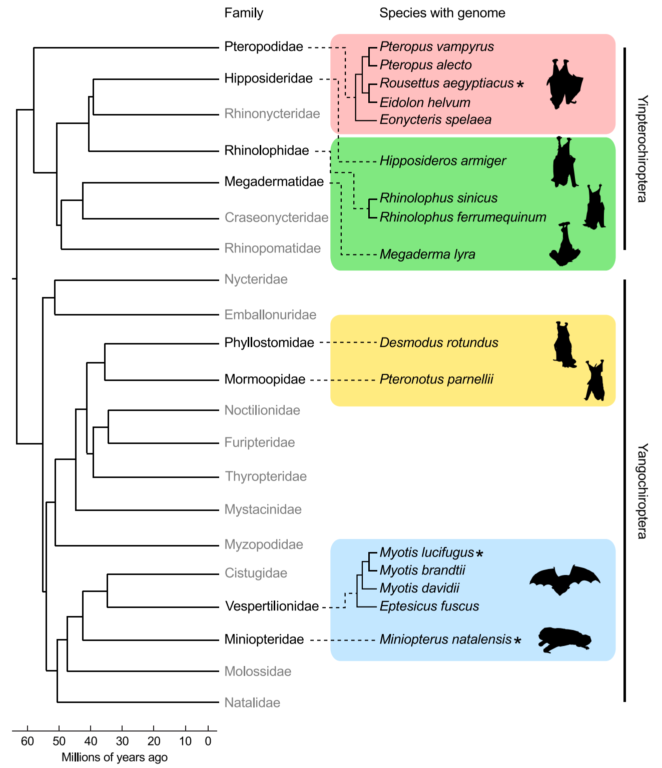


Figure 1. We used available genomes of 16 bat species from eight out of 21 families for non-coding RNA annotation in this study. The tree shows their phylogenetic relationship and is based on a molecular consensus on family relationships of bats (4), further adapted and extended from (2). Bat families and species with published genomes currently available in the NCBI are shown (details see Table 1). Bat families still lacking a published genome assembly are written in gray color. RNA-Seq data sets were selected from species marked with an asterisk and additionally obtained from a *Myotis daubentonii* cell line (see Table 2). Bat silhouettes were adapted from artworks created by Fiona Reid.

entific fields, including healthy ageing, immune and ecosystem functioning, the evolution of sensory perception and human communication, and mammalian genome architecture (see the recent Bat1K review for further details (2)).

Despite the unique biological characteristics of these flying mammals and their important role as natural reservoirs for viruses, bats are one of the least studied taxa of all mammalian (17). Accordingly, there is little knowledge about the non-protein-coding transcriptome of bats, which plays a crucial role in an extensive number of cellular and regular functions and comprises a very diverse family of untranslated RNA molecules (18,19). In addition, it is believed that due to the early evolutionary radiation of bats (compared to other mammals) their innate and acquired immune responses have a different set of molecules (1).

Genome assemblies and annotations are essential starting points for many molecular-driven and comparative studies (20). Especially, studies of non-model organisms play important roles in many investigations (21). In most cases, however, these organisms lack well-annotated genomes (22), which severely limit our ability to gain a deeper understanding of these species and may further impede biomedical research (23).

In this study, we comprehensively annotated non-coding RNAs in 16 available bat genome assemblies (Table 1). For each bat species, we provide final annotations that are compatible with current Ensembl and NCBI (National Center for Biotechnology Information) standards (GTF format) and that can be directly used in other studies, for example for differential gene expression analysis. We compare our new annotations with the currently available annotations for bats and show that a large number of non-coding genes are simply not annotated and are therefore overlooked by other studies. We used six RNA-Seq data sets comprising different conditions (Table 2) to validate our annotations and to determine the expression levels of our newly annotated ncRNAs. Exemplarily, we show that our novel annotations can be used to identify ncRNAs that are significantly differential expressed during viral infections and were missed by previous studies.

MATERIALS AND METHODS

Bat species and genomic assembly data

We downloaded the last recent genome versions for 16 bat species in September 2018 from the NCBI genome database (Table 1). Within the order of *Yinpterochiroptera*, nine genomic sequences were obtained covering the bat families *Pteropodidae*, *Hipposideridae*, *Rhinolophidae* and *Megadermatidae* whereas for the order of *Yangochiroptera* another seven genome assemblies were available for the bat families *Phyllostomidae*, *Mormoopidae*, *Vespertilionidae* and *Miniopteridae* (Figure 1). We introduced a unique three-letter abbreviation code (Table 1) to easily distinguish between the 16 bat species in the manuscript and intermediate annotation files provided in the Electronic Supplement. We used QUAST (v5.0.2) (24) to calculate several assembly statistics for all genomes, shown in Supplementary Table S1.

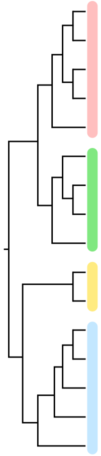
At the end of 2018, two new bat genomes were presented by the Bat1K project (http://bat1k.ucd.ie) (2), comprising a newer version of the greater horseshoe bat genome (*Rhinolophus ferrumequinum*; *Rhinolophidae*) and the genome of the pale spear-nose bat (*Phyllostomus discolor*; *Phyllostomidae*). However, these two bat genomes were not included in our present study due to the data use policy of the Bat1K consortium and to support a fair and productive use of these data.

RNA-Seq data

To validate our novel ncRNA predictions, we selected six RNA-Seq data sets (16,25–28) comprising all together 98 samples gathered from four different bat species. We have labeled each published RNA-Seq data set based on the first authors last name and the year of data set publication (Table 2 and Supplementary Table S2).

All samples were quality trimmed using Trimmomatic (29) (v0.36) with a 4 nt-step sliding-window approach (Q20) and a minimum remaining read length of 20 nt. For the *Field-2018* data set (28), we additionally removed the three leading 5' nucleotides from the reads of each sample because of a generally low quality observed by FastQC (www.bioinformatics.babraham.ac.uk/projects/

Table 1. We have annotated ncRNAs within 16 bat genomes of different assembly quality. We introduced three-letter abbreviations for each bat species used throughout the manuscript and in supplemental files and annotations. Genome sizes were estimated (*Est.*) by using *C*-values (DNA content per pg) from the animal genome size database (http://genomesize.com) and by applying the following formula: *Genome size* = $(0.978 \cdot 10^9) \cdot C$. If multiple entries for one species were available, an average over all *C*-values was calculated and used to estimate the genome size. If one species could not be found, an average *C*-value for the corresponding genus was used. Supplementary Table S1 provides additional assembly statistics calculated by QUAST (v5.0.2) (24). NCBI acc. – GenBank assembly accession without the prefix GCA_

	Species	Acronym	Size (Gb)	Est. (Gb)	# contigs	N50	GC (%)	NCBI acc.	Ref.	Year
	*Pteropus vampyrus*	PVA	2,198	2,318	36,094	5,954,017	39.81	000151845	(25)	2011
	Pteropus alecto	PAL	1,985	2,025	65,598	16,439,519	39.71	000325575	(26)	2013
	Rousettus aegyptiacus	RAE	1,910	2,064	2,490	2,007,187	40.02	001466805	(27)	2018
	Eidolon helvum	EHE	1,837	1,985	133,538	27,684	39.24	000465285	(28)	2013
	Eonycteris spelaea	ESP	1,966	2,181	4,469	8,002,591	40.30	003508835	(29)	2018
	Hipposideros armiger	HAR	2,236	2,516	7,571	2,328,177	41.11	001890085	(30)	2016
	Rhinolophus sinicus	RSI	2,073	2,136	63,439	3,791,306	40.51	001888835	(30)	2016
	Rhinolophus ferrumequinum	RFE	1,926	2,178	160,500	21,151	40.59	000465495	(28)	2013
	Megaderma lyra	MLY	1,735	–	192,872	16,881	40.39	000465345	(28)	2013
	Desmodus rotundus	DRO	2,063	2,519	29,801	26,869,735	42.14	002940915	(31)	2018
	Pteronotus parnellii	PPA	1,960	2,533	177,401	22,675	40.82	000465405	(28)	2013
	Myotis lucifugus	MLU	2,034	2,210	11,654	4,293,315	42.50	000147115	(25)	2011
	Myotis brandtii	MBR	2,107	2,255	169,750	3,256,877	42.55	000412655	(32)	2013
	Myotis davidii	MDA	2,059	2,255	101,769	3,493,173	42.69	000327345	(26)	2013
	Eptesicus fuscus	EFU	2,026	2,282	6,789	13,454,942	42.93	000308155	–	2012
	Miniopterus natalensis	MNA	1,803	1,933	1,269	4,315,193	42.06	001595765	(33)	2016

fastqc/) (v0.11.7). The remaining reads of all processed samples were individually mapped to all 16 bat assemblies using HISAT2 (30) (v2.1.0) and transcript abundances were subsequently calculated from all resulting 1568 mappings by featureCounts (31) (v1.6.3). If suitable, the appropriate strand-specific counting mode was applied for each data set (see Table 2 for information about the strand specificity). For each bat genome assembly, the merged annotation of already known (NCBI) and newly identified (this study) ncRNAs was used (Supplementary Files S1). Due to the size of the annotations and the huge amount of overlaps, long ncRNAs were counted and analyzed for differential expression separately.

To enable a better investigation of small RNAs (sRNAs), we included a data set of the targeted sequencing of sRNAs (especially miRNAs) from *M. daubentonii* cells (*Weber-2019*). To obtain this sRNA sequencing data, total RNA of 18 samples, which was obtained using the same procedure like explained for the rRNA-depleted *M. daubentonii* data (26), was preprocessed using the Illumina TruSeq smallRNA protocol, sequenced on one HiSeq 2500 lane, and finally uploaded in the course of this study under GEO accession GSE132336. The reads of these 18 sRNA samples were additionally preprocessed by removing potential adapter sequences with cutadapt (32) (v1.8.3) followed by a quality (Q20) trimming using again a window-size of 4 and a minimum length of 15 nt by PrinSeq (33) (v0.20.3). The processed sRNA samples were either individually mapped to the 16 bat genomes for differential expression analysis or combined and mapped on each bat genome to predict known and novel miRNAs with miRDeep2 (34).

Differential gene expression analysis

Only uniquely mapped reads were counted and used for the differential gene expression analyses with DESeq2 (35) (v1.16.1). Annotated rRNA genes were removed prior DESeq2 and TPM (transcripts per million) analysis. All raw read counts from samples of one data set were combined and normalized together using the built-in functionality of DESeq2, followed by pairwise comparisons to detect significant (adjusted p-value < 0.05; absolute $\log_2$ fold change $>$ 2) differential expressed ncRNAs.

Besides the DESeq2 normalization, we calculated TPM values for each ncRNA in each sample as previously described (36):

$$TPM_i = \frac{c_i}{l_i} \cdot \left(\frac{1}{\sum_{j \in N} \frac{c_j}{l_j}} \right) \cdot 10^6$$

where c_i is the raw read count of ncRNA i, l_i is the length of ncRNA i (and the cumulative exon length in the case of lncRNAs) and N is the number of all ncRNAs in the given annotation. To this end, we obtained for each RNA-Seq sample, each bat annotation, and each ncRNA one TPM value representing the normalized expression level of this ncRNA. If available, we calculated all TPM values in relation to the expression of all already known coding and non-coding genes and not only based on our novel ncRNA annotation.

Although we performed mappings, read countings, and normalization for all samples, bat genome assemblies and all six data sets (Table 2; overall 1568 mappings), we only selected one comparison per data set to exemplarily show novel and significantly differential expressed ncRNAs (Supplementary Files S2.1–S2.15; divided by data set and input annotation). For each data set, we chose the bat species that was also used in the corresponding study. For the *Hölzer-2019* and *Weber-2019* data sets, we used the closely related *M. lucifugus* genome assembly and annotation as a refer-

Table 2. Six RNA-Seq data sets comprising all together 98 samples derived from four different bat species were used to evaluate our novel ncRNA annotations. All samples were quality trimmed and individually mapped to all 16 bat assemblies using `HISAT2` (30) and transcript abundances were subsequently calculated from all 1568 mappings by `featureCounts` (31). We labeled each RNA-Seq data set based on the first authors last name and the year of data set publication. Raw read data of the enriched sequencing of small RNAs (especially miRNAs) of a *M. daubentonii* cell line (accompanying GSE121301 (26)) have been uploaded in the course of this publication under GEO accession GSE132336 (*Weber-2019*). polyA+ – library preparation with mRNA selection; rRNA- – library preparation with rRNA depletion and size selection (>200 nt); sRNA – library preparation with size selection (<200 nt); se/pe – single-/paired-end sequencing; ss/not-ss – strand-specific/unstranded sequencing; ss^s/ss^a – strand-specific in sense orientation/in antisense orientation

				Reads					
Nr.	Label	Bat species	Nr. samples	Length (bp)	Nr. (mio)	Library type	Seq. setup	Accession	Ref.
1	*Field-2015*	*M. lucifugus*	11	101	15,1–19,4	polyA+	pe/ss^a	SRP055976	(27)
2	*Eckalbar-2016*	*M. natalensis*	18	100	30,9–82,7	polyA+	pe/ss^s	SRP051253	(25)
3	*Hölzer-2016*	*R. aegyptiacus*	9	100	18,7–25,1	rRNA-	pe/not-ss	SRP128545	(16)
4	*Field-2018*	*M. lucifugus*	24	50	30,8–49,9	rRNA-	se/ss^s	SRP111376	(28)
5	*Hölzer-2019*	*M. daubentonii*	18	50	66,2–72,0	rRNA-	se/ss^a	GSE121301	(26)
6	*Weber-2019*	*M. daubentonii*	18	50	8,6–12,5	sRNA	se/ss^s	GSE132336	–

ence, because currently no genomic sequence for *M. daubentonii* is publicly available.

Field-2015 (11 samples, MLU). We compared control (*mock*) samples (five replicates) with the *infected* (white-nose syndrome, WNS) samples (six replicates) obtained from wing tissue of *M. lucifugus* to identify novel ncRNAs differentially expressed during the infection with the psychrophilic fungus *Pseudogymnoascus destructans* (27).

Eckalbar-2016 (18 samples, MNA). To identify new ncRNAs playing different roles during the development of the bat wing, we compared forelimb samples and hindlimb samples (each of three developmental stages in three replicates) of *M. natalensis* independent from the embryonic stages (25).

Hölzer-2016 (9 samples, RAE). Here, we investigated novel ncRNAs that might play a role due to transcriptional changes between uninfected (*mock*) samples and Ebola/Marburg virus infected samples regardless of the time point of infection and obtained from *R. aegyptiacus* cells (16). Due to the lack of real biological replicates, we only calculated TPM values and did not include this data set into the differential expression analysis with `DESeq2`.

Field-2018 (24 samples, MLU). As Field *et al.* were interested in host transcriptomic responses to a fungal pathogen (*P. destructans*, *Pd*) during torpor in hibernating bats (*M. lucifugus*), we searched for novel ncRNAs that were differentially affected between *P. destructans* positive (*Pd*+) samples, obtained although the bats were still torpid (3–6°C; six replicates), and *Pd*- samples taken after the bats were allowed to warm to euthermic temperature (six replicates) (28).

*Hölzer-2019***/***Weber-2019 (18 samples, MLU).* For these studies, RNA was extracted from 18 samples of *M. daubentonii* tissue either left uninfected (*mock*), infected with virus (RVFV Clone 13), or stimulated with interferon (IFN) at two different time points (26). Sequencing was performed with two different protocols resulting in data sets: *Hölzer-2019* (rRNA-depleted) and *Weber-2019* (smallRNA-enriched). For both data sets, we compared *mock* and virus-infected (Clone 13) samples at 24 h post infection (each with three replicates) in more detail (Supplementary Files S2) and present for all 18 smallRNA-Seq samples (*Weber-2019*) normalized expression values in Figure 4. Please note that for *M. daubentonii* currently no genome assembly is available, so the genome assembly of *M. lucifugus* was used as a close relative.

Annotation file format

All of our annotations follow the *General Transfer Format* (`GTF`) as described and defined in the `Ensembl` (37) database (https://ensembl.org/info/website/upload/gff.html). Therefore, each row of each annotation file is either defined as a *gene*, *transcript*, or *exon* (by the feature column) and strictly following a hierarchical structure, even if only one exon (as for most ncRNAs) is reported. By adhering to this annotation format, our novel annotations can be easily merged with existing ones as derived from `Ensembl` or `NCBI` and are directly usable as input for computational tools such as `HISAT2` for mapping or `featureCounts` for transcript abundance estimation. We defined gene, transcript, and exon IDs following the `Ensembl` pattern: <species><feature><ncRNA><11-digit-number>.
For example, the ID: MLUGR00000000001 denotes the first (00000000001) rRNA (R) gene (G) annotated in the *M. lucifugs* (MLU) genome. We defined the following abbreviations for different ncRNA types: rRNA (R), tRNA (T), miRNA (M), miRNA with `mirDeep2` (D), snoRNA (S), ncRNA/miscRNA/other (N), lncRNA (L), lncRNA hot-spot (H), mitochondrial ncRNA (O).

Annotation of non-coding RNAs

In general, we used specialized computational tools for the annotation of specific ncRNA classes (Supplementary Tables S3–S10). If not otherwise stated, the main ncRNA-discovery is based on homology searches against the `Rfam` database (38–40) (v14.0). We used the `Gorap` pipeline (https://github.com/koriege/gorap), a specially developed software suite for genome-wide ncRNA screening. `Gorap` screens genomic sequences for all ncRNAs present in the `Rfam` database using a generalized strategy by applying multiple filters and specialized software tools. To this end, `Gorap` takes huge advantage of `Infernal` (41,42) (v1.1.2) to annotate ncRNAs based on input alignment files conserved in sequence and secondary structure (so-called *Stockholm* alignment files; `stk`). All resulting alignment files were automatically pre-filtered by `Gorap` based on different ncRNA class-specific parameters including taxonomy, secondary structure and primary sequence comparisons. Due to repeats, pseudogenes, undiscriminable un-/functional genes and overlapping results from the different assembly methods, we defined a ncRNA set per species for annotation that includes filtered sequences, but allows for variants and multiple copies. This final annotation set is defined by hand-curating the resulting `stk` alignments of `Gorap` with the help of `Emacs RALEE` mode (43). Due to the removal of sequences in the stockholm alignments, the remaining sequences were extracted and again aligned into stockholm format using `cmalign --noprob` from `Infernal`. The `Rfam`-derived ncRNA alignments were further split into *snoRNAs* (Supplementary Table S5), *miRNAs* (Supplementary Table S6) and *other ncRNAs* including snRNAs, lncRNAs and other structural RNAs (Supplementary Table S8).

In general, our annotation results give an overview about the amount of different ncRNAs in bat species and intentionally can include false positive hits and duplicates. All ncRNA hits are placed as `STK` (if available), `GTF` and `FASTA`-files in the Electronic Supplement and OSF (doi.org/10.17605/OSF.IO/4CMDN).

rRNAs. We used the prediction tool `RNAmmer` (44) (v1.2) to identify 5.8S, 18S and 28S rRNA genes using hidden markov models. The tools output was transformed into regular `GTF` file format. All output files can be found in Supplementary Table S3.

tRNAs. For the annotation of tRNAs, we applied `tRNAscan-SE` (45) (v1.3.1) to the bat contigs using default parameters. The results were filtered by removing any tRNAs of type 'Undet' or 'Pseudo' and the tabular output was transformed into the `GTF` file format. Additional information about the anticodon and the coverage score were added to the description column. We provide the raw `tRNAscan-SE` files in Supplementary Table S4.

snoRNAs. We annotated snoRNAs based on available alignments from the `Rfam` database using `Gorap` and additionally marked and classified them into box C/D and box H/ACA when intersecting with the set of snoRNAs from http://www.bioinf.uni-leipzig.de/publications/supplements/12-022 (46) (Supplementary Table S5).

miRNAs. Additionally to the `Rfam`-screening (Supplementary Table S6), miRNAs were predicted by the `miRDeep2` pipeline (34) (v2.0.0.8) using default parameters and the combined smallRNA-Seq data set (*Weber-2019*; 18 samples) mapped to each individual bat assembly as an input for `miRDeep2` (Supplementary Table S7).

To validate the accuracy of our approach, we compared our `miRDeep2` annotations of *M. lucifugus*/*P. alecto* (based on the transcriptomic data derived from *M. daubentonii*; *Weber-2019*) with annotations of miRNAs for transcriptomic data of *Myotis myotis* (8) and *P. alecto* (47). For reference mapping, Huang *et al.* also used the *M. lucifugus* genome, so we were theoretically able to directly compare our annotations with the annotations of both studies. Unfortunately, no positional information (annotation file) of the identified miRNAs derived from the transcriptomic data of *M. myotis* were given in the manuscript or supplement (8). Therefore, we blasted the precursor miRNA sequences identified with the help of the *M. myotis* transcriptome against the *M. lucifugus* genome and retained only hits with a sequence identity of 100%. The so obtained positional information was further used to calculate the overlap between our predicted miRNAs in *M. lucifugus*. We used the same approach for the *P. alecto* comparison. If the annotated location of an miRNA and one of our identified miRNA locations in *M. lucifugus*/*P.alecto* were overlapping by at least 85 %, we counted this location as a common prediction.

lncRNAs. Long ncRNAs were annotated using a high confidence data set H from the `LNCipedia` (48) (v5.2) database comprising 107 039 transcripts of potential human lncRNAs. The transcripts were used as input for a `BLASTn` (2.7.1+, $1e^{-10}$) search against each of the 16 bat assemblies (compiled as `BLAST` databases). The `BLASTn` result for each bat assembly was further processed to group single hits into potential transcripts as follows: first, for each query sequence $q \in H$, hits of q found on the same contig c and strand s were selected ($hits_{c,\,s,\,q}$) and the longest one, q_1, was chosen as a starting point so that $trscp_{c,\,s,\,q} = (q_1)$. Second, all hits $q_i \in hits_{c,\,s,\,q}$ with $q_i \notin trscp_{c,\,s,\,q}$, which overlap neither in the query q nor in the target sequence and do not exceed a maximum range of 500 000 nt from the most upstream to the most down-stream target sequence position of all $q_j \in trscp_{c,\,s,\,q} \cup q_i$, were added iteratively to $trscp_{c,\,s,\,q}$. To this end, we introduced a simple model of exon–intron structures, naturally occurring when using transcript sequences as queries against a target genome assembly. We defined the 500 000 nt search range based on an estimation of lncRNA gene sizes derived from the human `Ensembl` (37) annotation. If the sum of the lengths of all $q_i \in trscp_{c,\,s,\,q}$ covers the query transcript length $length_q$ at least for 70 %, $trscp_{c,\,s,\,q}$ was considered as a transcript and its elements q_i as exons, otherwise all $q_i \in trscp_{c,\,s,\,q}$ were withdrawn. This procedure is repeated until all hits $\in hits_{c,\,s,\,q}$ were used or withdrawn. Therefore, each so-defined group of non-overlapping hits derived from the same query sequence and found on the same contig and strand should represent a lncRNA transcript with its (rough) exon structure. The defined transcripts were saved as `BLAST`-like output and transformed into `GTF` file format. To follow the `GTF` annotation format and to harmonize our lncRNA annotations with the other ncRNA annotations, each lncRNA transcript was also saved as a gene annotation and consists of at least one exon.

As we observed a lot of different sequences from `LNCipedia` aligning to the same positions in the genomes, we decided to condense exons at the same sequence positions, considering transcripts with one or multiple exons separately. For each contig and strand, starting from the 5' end, exons with a minimum overlap of 10 nt were grouped together. In the case of multiple exons, groups of exons were merged, if they shared any transcript origin. If all exons in the group originated from the same LNCipedia gene, the group was considered as one gene with several transcripts and its associated exon(s). Otherwise, we defined a lncRNA *hot spot* on gene level with several transcripts and their associated exon(s). The LNCipedia names of the gathered transcripts of a lncRNA hot spot, as well as start and end positions of all exons, are listed in the GTF gene attribute field (Supplementary Table S9). The scripts used for the identification of lncRNAs can be found at https://github.com/rnajena/bats_ncrna.

Mitochondrial DNA. As not all of the 16 genome assemblies are including contigs representing the mitochondrial DNA (mtDNA), we downloaded for MLU (KP273591), MBR (KM199849), MDA (KM233172), PAL (NC_023122), PVA (KP214033), PPA (KF752590), RFE (NC_020326), RAE (NC_007393), HAR (NC_018540), and DRO (NC_022423) mitochondrial genomes from the NCBI (Table 3). For the other six species we used BLASTn (2.7.1+, $1e^{-10}$) with the MLU and PVA mitochondrial genomes as queries against the remaining bat genomes. For EHE, we found a possible mtDNA contig in full-length (16 141 nt; AWHC01200796) in the genome assembly. Due to the circularization of mtDNA, we rearranged the sequence of this contig to start with the gene coding for the phenylalanin tRNA and to match the gene order of the other mitochondrial genomes. Only for ESP, RSI, MLY, EFU and MNA, we were not able to detect any possible contigs of mtDNA (Table 3). All 11 mitochondrial genomes were annotated with MITOS (49). The ncRNA results were filtered by e-value (threshold 0.001), thus one of two small rRNAs in MBR and RFE and one of two large rRNAs in EHE were discarded as false positive hits (Supplementary Table S10). For the five bat assemblies directly including mtDNAs (Table 3), the MITOS annotations were added to the final merged ncRNA annotation. All other mtDNAs and annotations can be found in Supplementary Table S10.

Table 3. Mitochondrial bat genomes (mtDNA) publicly available and used for annotation with MITOS (49). For 10 out of the 16 bat species investigated in this study, mtDNA could be found in the NCBI. For four bat species, the mtDNA is also part of the genome assembly as determined using BLASTn. For *E. helvum*, no mtDNA could be found in the NCBI, but we were able to identify a single contig that is part of the genome assembly as mtDNA using BLASTn and the mitochondrial genomes of the other bats as query. The contig was rearranged to match the gene order of the other mtDNAs. ^{R} – found via BLASTn and rearranged

Species	Contig ID in assembly	NCBI accession
PVA	–	KP214033
PAL	NC_023122	NC_023122
RAE	NC_007393	NC_007393
EHE^R	AWHC01200796	–
ESP	–	–
HAR	NC_018540	NC_018540
RSI	–	–
RFE	–	NC_020326
MLY	–	–
DRO	NC_022423	NC_022423
PPA	–	KF752590
MLU	–	KP273591
MBR	–	KM199849
MDA	–	KM233172
EFU	–	–
MNA	–	–

Computational merging of gene annotations

As we annotated all bat assemblies by using different tools, we needed to merge the resulting GTF files to resolve overlapping annotations and to receive a final annotation of ncRNAs for each bat species. Furthermore, we extended the available NCBI annotations (including protein- and non-coding genes) by integrating our novel ncRNA annotations. The scripts used to merge the different annotation files and to calculate overlaps between annotations can be found at https://github.com/rnajena/bats_ncrna. Due to their size, we have not included the lncRNA annotations based on LNCipedia. These can be downloaded and used separately (Supplementary Table S9).

Merge of novel non-coding annotations. For each bat species, we merged the ncRNA annotations (except for lncRNAs) using a custom script (merge_gtf_global_ids.py). After reading in all features and asserting correct file structure, overlaps were resolved in the following manner: (i) Exons are considered overlapping if >50 % of the shorter one is covered by the larger one. (ii) If only one of the overlapping set is of biotype *protein-coding*, remove all others. (iii) For further ties, keep only the exon that is highest on a priority list based on annotation source. (iv) For further ties, keep only the longest of the exons. For each exon to be removed the corresponding transcript is deleted, and gene records that lost all transcripts are also deleted.

Merge of NCBI and novel annotations. We first converted and filtered the NCBI annotations to a compatible format with a custom script (format_ncbi.py) and then combined the results with our merged novel ncRNA annotations using the same strategy to resolve overlaps as above, but imposing less strict format rules (merge_gtf_ncbi.py).

RESULTS

Assembly and annotation quality differs among bat species

At best, a genome assembly represents the full genetic content of a species at chromosome level. Whereas the first complex eukaryotic genomes were generated using Sanger chemistry, today's technologies such as Illumina short-read sequencing and PacBio or Oxford Nanopore long-read approaches are increasingly used (50). The currently available bat genomes vary widely regarding their assembly quality and completeness (Table 1 and Figure 2 ; Supplementary Table S1) and were predominantly assembled by using short Illumina-derived reads and low (∼18 X) (51) up to moderate/higher coverage (77–218 X) approaches (52–56,25). A new assembly of the cave nectar bat (*Eonycteris spelaea*) (57) was exclusively generated from long-read data derived from the PacBio platform, and the genome of the Egyptian fruit bat (*Rousettus aegyptiacus*) (58) was assembled by using a hybrid-approach of Illumina and PacBio

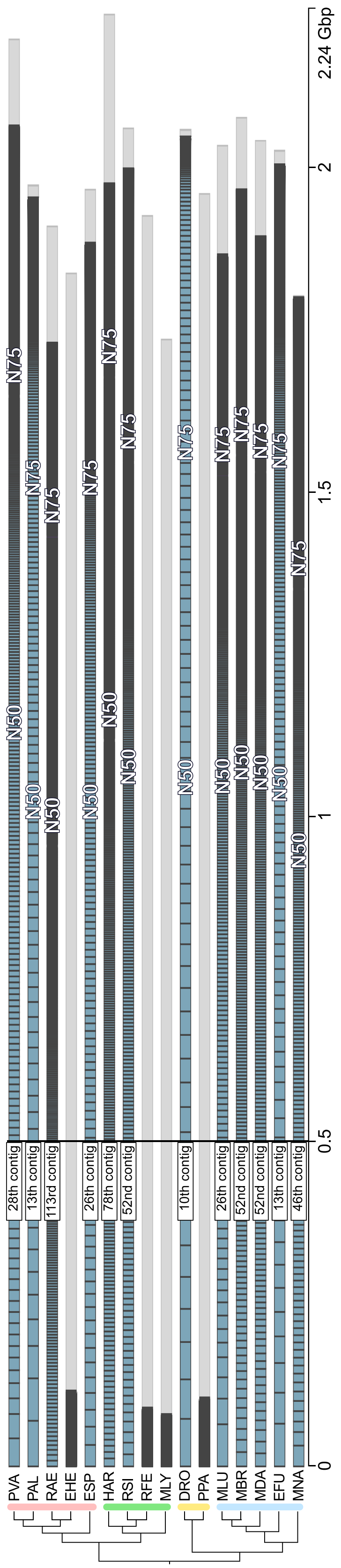


Figure 2. A contig size view of all bat genome assemblies annotated in this study. Only the largest 1000 contigs of each assembly are shown. Gray bars indicate all contigs that are shorter than the 1000 largest contigs. N50/N75 labels mark the position of the contig with length equal to the N50/N75 value. Due to a low-coverage approach (51), the assemblies of *E. helvum*, *R. ferrumequinum*, *M. lyra* and *P. parnellii* are of generally lower quality (see Table 1 and Supplementary Table S1). Thus, N50/N75 values are not visualized for these species. The figure was adapted from the `Icarus` viewer (70) used as a part of `QUAST` (24).

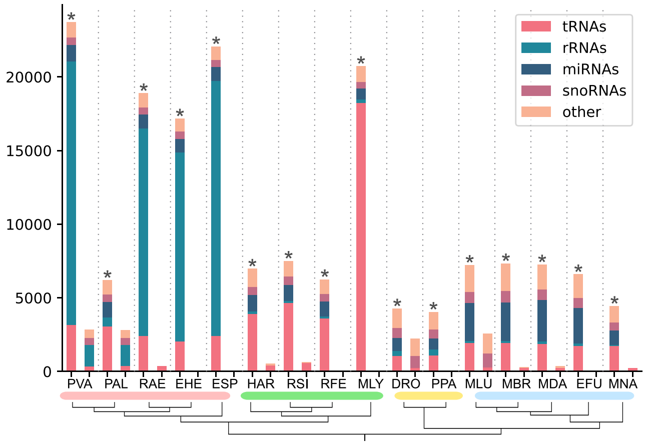


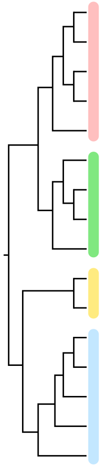
Figure 3. The number of already annotated ncRNAs (`NCBI`) and newly annotated ncRNAs in this study (marked by an asterisk) for each bat species. Due to the general detection approach (see Methods section) and their high number lncRNAs are omitted from this figure. Details about lncRNAs can be found in Supplementary Table S9.

data. These two genomes from the *Pteropodidae* family are of a generally higher quality (Figure 2 and Supplementary Table S1). Regardless of their assembly quality, these genomes need to be annotated to identify regions of interest, for example, encoding for protein- and non-coding genes or other regulatory elements.

Non-coding RNAs are underrepresented in current bat genome annotations

Current genome annotations, mostly generated by automatic annotation pipelines provided by databases such as the `NCBI` (59) or `Ensembl` (37), are predominantly focusing on protein-coding genes and well-studied ncRNAs such as tRNAs and rRNAs. Accordingly, the available bat genome annotations vary a lot regarding their quality, ranging from more comprehensive annotations for long-standing bat genomes such as *M. lucifugus* or *P. vampyrus* to annotations on *region* level, completely missing any coding or non-coding gene annotations at the current `NCBI` version (Figure 3 and Table 4). Furthermore, by using strand-specific RNA-Seq data, we could show that some genes (e.g. *IFNA5*/*IFNW2* in the `Ensembl` annotation of *M. lucifugus* (26)) are annotated on the wrong strand and are therefore entirely missed by differential expression studies when relying on a strand-specific read quantification. For all publicly available bat genomes, ncRNAs are generally annotated on low levels and are highly incomplete, mostly only comprising some tRNAs, rRNAs, snRNAs, snoRNAs and lncRNAs (Figure 3 and Table 4). Therefore, many ncRNAs, especially miRNAs, are simply overlooked by current molecular studies, for example from RNA-Seq studies that aim to call differential expressed genes based on such in-complete genome annotation files. Studies that have made additional effort on annotating ncRNAs in bats (8,25,47,60–61) are not reporting their results on a level that can be directly used for further computational assessment (e.g. as a direct input for RNA-Seq abundance estimation).

Table 4. Currently annotated protein-coding genes and ncRNAs of the 16 bat genome assemblies obtained from `NCBI`. Genes were counted by checking the *gene biotype* tag for all *gene* entries (third column in the `GTF` file). Five bat species (EHE, ESP, RFE, MLY, PPA) are currently missing any protein- and non-coding gene annotations and only provide *regions* for each scaffold in the assembly. The annotation of *E. fuscus* is completely missing any *gene biotype* tags. For the description of the three-letter abbreviations, please refer to Table 1. misc. RNA – miscellaneous RNA, not classifying into the other groups



Species	proteins	tRNA	rRNA	miRNA	snoRNA	snRNA	lncRNA	misc. RNA
PVA	19,529	360	1,447	0	488	523	3,251	11
PAL	18,363	370	1,433	0	479	506	2,250	9
RAE	19,425	340	2	0	0	0	2,333	0
EHE	0	0	0	0	0	0	0	0
ESP	0	0	0	0	0	0	0	0
HAR	18,962	413	2	0	0	0	2,532	113
RSI	21,417	597	0	0	0	0	2,035	56
RFE	0	0	0	0	0	0	0	0
MLY	0	0	0	0	0	0	0	0
DRO	18,631	237	4	0	824	1,094	402	0
PPA	0	0	0	0	0	0	0	0
MLU	19,953	262	32	0	926	1,258	2,192	15
MBR	19,988	250	0	0	0	0	4,083	42
MDA	18,851	263	0	0	0	0	3,141	99
EFU	24,147	238	0	0	0	0	0	2
MNA	18,161	223	0	0	0	0	1,311	0

Currently, in the `NCBI` database, five bat assemblies are entirely lacking any coding/non-coding annotations and miRNAs are not annotated at all (Table 4). The `Rfam` database (39) contains mainly for *M. lucifugus* and *P. vampyrus* 336 ncRNA families. Other ncRNAs are currently unknown from bat genomes or not well documented.

Identification and validation of ncRNAs in 16 bat genomes

The genome assembly status of different bat species varies widely: ranging for example from 29 801 contigs and an N50 of 26 869 735 nt (*D. rotundus*) to 192 872 contigs and an N50 of only 16 881 nt (*M. lyra*), see Table 1 and Supplementary Table S1. Accordingly, the annotation status also varies a lot (Table 4). Within this work, we give an overview of potential ncRNA annotations in bats. However, the precise number of ncRNAs remains unclear, because of ncRNAs being present several times in the assemblies, and others still remaining undiscovered.

To give a better estimation of transcribed and potentially functional ncRNAs, we used six Illumina short-read RNA-Seq data sets derived from four bat species (Table 2) to estimate the expression levels of our novel annotations. Note that we refer throughout this paper to an RNA-Seq data set by the first author's name and the year of the respective data set publication. The only included data set derived from a species of the *Yinpterochiroptera* suborder (*R. aegyptiacus*) was obtained from a study dealing with the differential transcriptional responses of Ebola and Marburg virus infections in human and bat cells (16) (data set: *Hölzer-2016*). In this study, total RNA of nine samples of R06E-J cells, either challenged by the Ebola or Marburg virus or left un-infected, were harvested at 3, 7 or 23 h post infection (poi) and sequenced. Unfortunately, no biological replicates could be generated for this study. Therefore, we did not use this data set for the differential expression analysis, but also calculated normalized expression values (TPM; transcripts per million) as done for all RNA-Seq data sets. The other five data sets comprise *Yangochiroptera* species of the *Vespertilionidae* (*M. lucifugus, M. daubentonii*) and *Miniopteridae* (*M. natalensis*) families (Table 2). Field *et al.* conducted two transcriptomic studies (27,28) (*Field-2015*, *Field-2018*) using wing tissue of the hibernating little brown myotis bat. They were especially interested in transcriptional changes between un-infected wing tissue and adjacent tissue infected with *Pseudogymnoascus destructans*, the fungal pathogen that causes the white-nose syndrome. Two other data sets were obtained from virus- (RVFV Clone 13) and interferon (IFN) alpha-induced transcriptomes of a *Myotis daubentonii* kidney cell line (MyDauNi/2c). RNA of mock, IFN and Clone 13 samples were gathered at two time points, 6 and 24 h poi (26). From the same samples, rRNA-depleted (*Hölzer-2019*) and smallRNA-concentrated (*Weber-2019*; see Methods section for details) libraries were generated and sequenced. Finally, we used data of the long-fingered bat *M. natalensis* initially obtained to characterize the developing bat wing (25). Here, total RNA was extracted from paired forelimbs and hindlimbs from three individuals at three developmental stage.

We have mapped each sample to each bat genome, regardless of the origin of the RNA. Expectedly, the mapping rate decreases in bat species that are evolutionary more far away from the original species from which the RNA was sequenced. However, we were interested to find out which ncRNAs are consistently transcribed in all investigated bat species or only in certain bat families and sub-groups.

Over all 16 bat assemblies, we annotated ncRNA families for in total 23 tRNAs, 3 rRNAs, 193 miRNAs (2680 predicted by `mirDeep2`), 162 snoRNAs, 22 mitochondrial (mt-)tRNAs and 2 mt-rRNAs as well as 244 other ncRNAs additionally derived from the `Rfam` database (39) (selected ncRNAs are shown in Table 5). With a broad approach, we have identified 24 316 potential lncRNAs and defined between 27 149 (*M. lyra*) and 158 135 (*D. rotundus*) lncRNA hot spots.

All annotations, separated for each ncRNA class and summarized for each bat species, can be found in the Open Science Framework (doi.org/10.17605/OSF.IO/4CMDN) and in our Electronic Supplement (rna.uni-jena.de/supplements/bats) and are compatible with the genome assembly versions listed in Table 1. Thus, our annotations together with the bat genome assemblies obtained from the NCBI can be directly used for subsequent analysis such as differential gene expression detection.

rRNAs

We detected 18S and 28S rRNA for the majority of investigated bat species (Supplementary Table S3). The varying number of rRNAs is in line with all currently available metazoan genomes, lacking the correct composition of rRNAs due to misassemblies. However, the number of 5.8S rRNA varies a lot between 17 919 for *P. vampyrus* and 51 for *M. natalensis* (Table 5 and Supplementary Table S3). Interestingly, 4 of 5 *Pteropodidae* show a higher number of 5.8S rRNAs compared to the other species (∼50–100 fold). Only the *P. alecto* assembly is in line with the other bat assemblies in regard to the amount of 5.8S rRNA copies.

tRNAs

For all bat species, we observed various numbers of tRNAs (Table 5 and Supplementary Table S4). We could detect full sets of tRNAs for *E. fuscus* and *M. davidii*, whereas between one and four tRNAs are missing from the assemblies of *H. armiger*, *M. brandtii*, *M. lucifugus*, *M. lyra*, *M. natalensis*, *P. parnellii*, *R. ferrumequinum*, *R. sinicus* and *D. rotundus,* and between nine and twelve are missing for *E. helvum*, *P. alecto*, *P. vampyrus*, *R. aegyptiacus* and *E. spelaea* (Supplementary Table S4). Interestingly, we identified a large number of tRNAs (18 235) for *M. lyra* in comparison to all other bat species (Supplementary Table S4), likely a result of the low assembly quality (Figure 2 and Supplementary Table S1). The lowest amount of tRNAs was annotated for *P. parnellii* and *D. rotundus* with only 1059 and 1041 copies, respectively. In all other bat genomes, we found between 1735 and 4657 tRNA genes (Supplementary Table S4).

The tRNA encoding for valine (Val) with the anticodon structure TAC had high copy numbers (over 1000) in all genome assemblies of the *Pteropodidae* family (Table 5). Similar copy numbers were achieved by the *R. ferrumequinum* and *R. sinicus* assemblies regarding the tRNA encoding for isoleucine (Ile) with the anticodon AAT (Supplementary Table S4). For tRNA(Ile) and the anticodon GAT, we also observed high copy numbers in *H. armiger*. Interestingly, all species with high tRNA(Val) and tRNA(Ile) copy numbers had rather low counts (between 0 and 28) of tRNA(SeC) with the anticodon TCA, while this tRNA was found with higher copy numbers in *P. parnelli* (68) and in *D. rotundus* (145) and with even higher counts (between 315 and 498) in all other bat species (Supplementary Table S4). However, high copy numbers might be also occur due to assembly quality and false positive predictions of tRNAscan-SE.

snoRNAs

In Supplementary Table S5 we list all detected snoRNAs, divided into box C/D and box H/ACA types. Overall, we found 162 snoRNA families within the investigated bat species, comprising 88 box C/D, 61 box H/ACA and 13 unclassified snoRNAs.

Many snoRNAs were found with exactly one copy present in each bat genome assembly (e.g. SCARNA7), whereas others were found in multiple copies for each bat species (e.g. SNORD112) or completely absent from certain bat families (e.g. ACA64), see Table 5. Exactly one copy of the small nucleolar RNA ACA64 was found within the genomes of the *Pteropodidae* family and multiple copies for *D. rotundus*, *P. parnellii*, *M. natalensis* and members of the *Vespertilionidae* family; however, this snoRNA seems to be completely absent from bat species of the *Megadermatidae* and *Rhinolophidae* families (Table 5). The H/ACA box snoRNA ACA64 is predicted to guide the pseudouridylation of 28S rRNA U4331 (62). Interestingly and as another example, SNORA48 was found in higher copy numbers in the genomes of the *Vespertilionidae* family. Among others, this H/ACA box snoRNA was described to be commonly altered in human disease (63).

miRNAs

Over all bat assemblies, we detected 193 miRNA families based on Rfam alignments (Supplementary Table S6) and predicted between 349 (*E. helvum*) and 2464 (*M. davidii*) miRNAs based on the 18 combined small RNA-Seq data sets (*Weber-2019*) using miRDeep2 (34) (Supplementary Table S7). The higher number of miRNAs predicted for *Myotis* species can be explained because the small RNA-Seq data set is derived from a *Myotis daubentonii* cell line.

Similar to other ncRNA classes, we observe various differences in miRNA copy numbers between the bat families. For example, mir-454 and mir-563 are absent in all *Vespertilionidae* and *M. natalensis* (Table 5), whereas mir-1912 is present in all *Yangochiroptera* (except *D. rotundus*) but absent from all *Pteropodidae* (Supplementary Table S6). The miRNA 541 is absent from all *Yangochiroptera* except *D. rotundus*. There are many other examples of absent/present miRNAs in certain bat species/families such as mir-662 (absent from *Pteropodidae*), mir-767 (absent from *Yinpterochiroptera*) and mir-626 (absent from *Rhinolophidae* and *Megaderma lyra*), see Supplementary Table S6 .

With miRDeep2, we detected hundreds of potential miRNAs for all investigated bat species (Supplementary Table S7). Generally, all 16 bat species can be divided into two groups. For the majority of the bat assemblies (12 out of 16) about 400 miRNAs were predicted. For the other 4 species of the *Vespertilionidae* family about 5 times as many miRNAs could be found. This is in concordance with the small RNA-Seq data used for the prediction, that was obtained from *Myotis daubentonii* kidney cells (see Methods). Nevertheless, ∼400 conserved miRNAs can be predicted in the evolutionary more distant bat species.

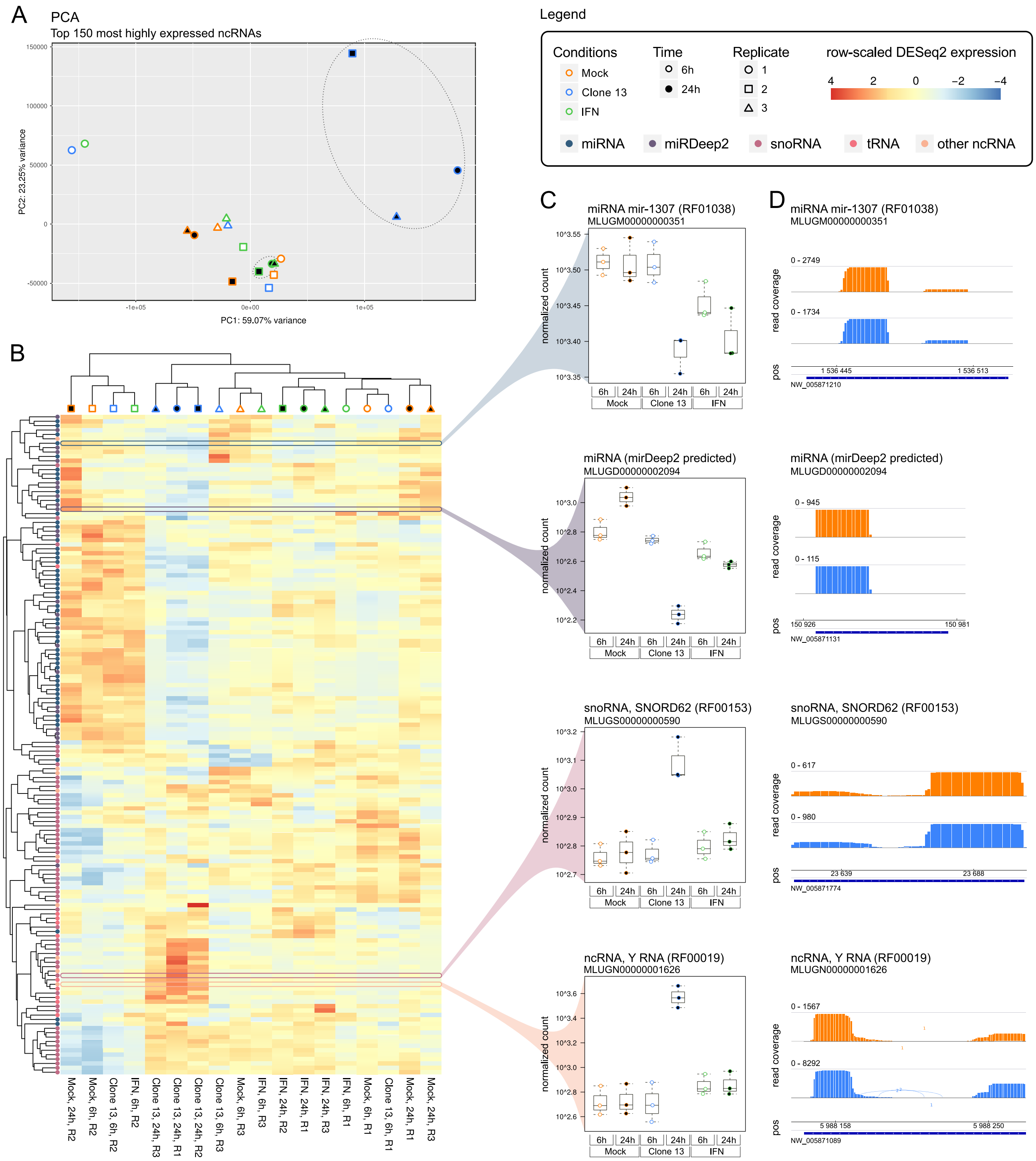
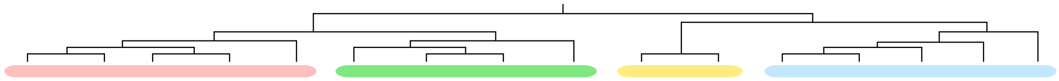


Figure 4. Results shown here are based on expression patterns obtained from the small RNA-Seq samples (*Weber-2019*) mapped to the *M. lucifugus* reference genome (NCBI: GCA_000147115). (**A**) A PCA performed by PCAGO (https://github.com/hoelzer-lab/pcago) using the 150 most highly expressed ncRNAs (75 miRNAs, 56 snoRNAs, 10 tRNAs, 9 other small ncRNAs; discarding rRNAs) as input. As also shown in the heat map (B), replicates tend to cluster on the basis of their passaging history, except for virus- and interferon-treated (IFN) samples 24 h post infection. (**B**) A heat map and hierarchical clustering for the same 150 most highly expressed ncRNAs that were used for PCA (A). The clustering was performed on DESeq2-normalized read counts and scaled on rows to visualize changes in expression on ncRNA level. Most differences can be observed between the Clone 13-infected samples 24 h post infection and all other conditions. Due to minor changes in miRNA expression levels, the other samples tend to cluster on the basis of their passaging history and not on the basis of their treatment. In general, miRNAs tend to be down-regulated (upper half), while snoRNAs tend to be up-regulated (lower half) after 24 h of Clone 13 infection. (**C**) Selected expression box plots of four highly expressed ncRNAs. Shown are DESeq2-normalized expression counts. The miRNA mir-1307 and a novel mirDeep2-predicted miRNA are significantly (adjusted p-value < 0.05) down-regulated at 24 h post virus infection. The C/D box snoRNA SNORD62 and a Y RNA are up-regulated. (**D**) IGV-derived expression patterns and raw read counts of the four selected ncRNAs obtained from one replicate of the mock and Clone 13 24 h conditions. Blue bars indicate the annotation and strandness.

Table 5. General genome information for each of the 16 investigated bat assemblies and selected ncRNA examples annotated in this study. We selected ncRNAs with interesting copy number distributions among the investigated bat species. For SNORA38, PVT1_5 and HOTTIP_2 and 3 we additionally found sophisticated differential expression patterns in at least one of the used RNA-Seq data sets (absolute log2 fold-change greater 1, TPM >10). Full tables and detailed information for each ncRNA class (FASTA, STK, GTF files) can be found in the Electronic Supplement online (Supplementary Tables S3-S10). Members of the *Vespertilionidae* as well as *D. rotundus* and *M. natalensis* appear to have a slightly higher GC-content (+ ∼2 %). PVA – *Pteropus vampyrus*; PAL – *Pteropus alecto*; RAE – *Rousettus aegyptiacus*; EHE – *Eidolon helvum*; ESP – *Eonycteris spelaea*; HAR – *Hipposideros armiger*; RSI – *Rhinolophus sinicus*; RFE – *Rhinolophus ferrumequinum*; MLY – *Megaderma lyra*; DRO – *Desmodus rotundus*; PPA – *Pteronotus parnellii*; MLU – *Myotis lucifugus*; MBR – *Myotis brandtii*; MDA – *Myotis davidii*; EFU – *Eptesicus fuscus*; MNA – *Miniopterus natalensis*. ACA64 – RF01225; SNORD112 –RF01169; SCARNA7 – RF01295; SNORA61 – RF00420; SNORA48 – RF00554; SNORA38 – RF00428; mir-454 – RF00746; mir-563 – RF01003; mir-662 – RF00983; mir-626 – RF00968; ZNRD1-AS1_2 – RF02219; HOXA11-AS1_1 – RF02137; LINC00901 – RF01884; PVT1_5 – RF02168; HOTTIP_2, 3 – RF02041, RF02042; U11 – RF00548; U4atac – RF00618; Hammerhead_HH10, HH9 – RF02277, RF02275; G-CSF_SLDE – RF00183



Species	PVA	PAL	RAE	EHE	ESP	HAR	RSI	RFE	MLY	DRO	PPA	MLU	MBR	MDA	EFU	MNA
General genome features																
Contigs	36,094	65,598	2,490	133,538	4,469	7,571	63,439	160,500	192,872	29,801	177,401	11,654	169,750	101,	6,789	1,269
Genome size (Gb)																
Assembly	2,198	1,985	1,910	1,837	1,966	2,236	2,073	1,926	1,735	2,063	1,960	2,034	2,107	2,059	2,026	1,803
AGSD[a]	2,318	2,025	2,064	1,985	2,181	2,516	2,136	2,178	–	2,519	2,533	2,210	2,255	2,255	2,282	1,933
GC-content (%)	39.81	39.71	40.02	39.24	40.30	41.11	40.51	40.59	40.39	42.14	40.82	42.50	42.55	42.69	42.93	42.06
ncRNA annotation																
rRNA families																
5.8S	17919	585	14038	12853	17278	194	128	106	259	351	393	146	142	162	169	51
tRNA families																
Val(TAC)	2405	2347	1764	1400	1788	4	5	4	8	4	8	7	5	7	5	5
Leu (GAG)	0	0	0	0	0	16	15	6	6	8	5	12	18	26	15	28
snoRNA families																
ACA64	1	1	1	1	1	0	0	0	0	2	3	3	2	1	3	2
SNORD112	60	68	57	57	41	60	73	72	21	29	37	30	36	25	29	40
SCARNA7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SNORA61	4	3	2	3	2	1	2	2	2	15	8	11	11	12	12	1
SNORA48	3	3	2	4	5	5	4	4	4	7	8	65	60	67	65	13
SNORA38	3	4	4	4	3	3	3	4	2	11	10	11	13	12	19	6
miRNA families																
mir-454/mir-563	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
mir-662	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
mir-626	1	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1
lncRNA families[b]																
ZNRD1-AS1_2	2	2	2	3	2	3	4	3	2	0	1	1	2	1	1	2
HOXA11-AS1_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
LINC00901	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PVT1_5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
HOTTIP_2, 3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
other families[c]																
U11	4	4	4	3	4	6	3	2	4	10	11	23	24	21	15	17
U4atac	1	1	1	1	1	2	2	3	1	4	6	3	4	3	6	3
Hammerhead_HH10, HH9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
G-CSF_SLDE	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

[a] estimated genome size based on C-values (DNA content per pg) obtained from the *Animal Genome Size Database* (for multiple entries an average over all C-values was calculated);
[b] long ncRNAs based on Rfam; [c] including remaining ncRNA families from the Rfam database, such as snRNAs and other unclassified regulatory ncRNAs.

We compared our miRDeep2-predictions in *M. lucifugus* and *P. alecto* with the predictions of two other studies (8,47). From the 540 published miRNAs based on the transcriptome of *M. myotis* (8) and the 426 published miRNAs based on transcriptomic data of *P. alecto*, we were able to obtain 490 and 368 miRNAs with positional information using BLASTn against the *M. lucifugus* and *P. alecto* genomes, respectively. From these 490 and 368 miRNAs, our prediction, based on the transcriptome of *M. daubentonii* (*Weber-2019*) mapped to *M. lucifugus* and *P. alecto*, included 195/490 (39.8 %) and 182/368 (49.5 %) miRNAs.

lncRNAs

For the annotation of lncRNAs, we have deliberately chosen a broad Blast-based approach, using 107,039 transcripts of potential human lncRNAs obtained from the LNCipedia database (48).

We have consciously chosen this approach, because lncRNAs have diverse genomic contexts, reveal various functions and act in different biological mechanisms (64–66). From 107 039 LNCipedia transcripts, we annotated 24 316 genes and 182 451 lncRNA *hot spots*. We defined regions in a genome as a lncRNA hot spot, if different LNCipedia transcripts derived from different genes map to the same region (see Methods section for detailed description). Overall, we found between 58 425 and 137 161 potential lncRNA genes in *M. lucifugus* and *R. sinicus*, respectively, and between 27 149 and 158 135 lncRNA hot spots in *M. lucifugus* and *D. rotundus*, respectively. We annotated the previously described lncRNAs Tbx5-as1 and Hottip (25) in all bat genomes, except Tbx5-as1 in *M. lucifugus*, presumably due to the lower genome assembly quality (Table 1).

Besides evolutionarily explainable differences in the presence of lncRNAs, we have observed that the number of lncRNAs and lncRNA hot spots increases with increasing assembly quality (e.g. with a higher N50; see Supplementary

Figure S1). We have not observed such a clear correlation between assembly quality and annotation results for short ncRNAs.

Other ncRNA elements

Based on the Rfam alignments, we were able to detect 244 other ncRNA families in addition to the rRNAs, tRNAs, snoRNAs and miRNAs described before (Supplementary Table S8). Overlaps with annotated lncRNAs (Supplementary Table S9) are intentional, because the Rfam includes only highly structured parts of long ncRNAs.

The highest number of ncRNA copies (951) was detected for the U6 spliceosomal RNA in *M. brandtii*, already known to have a lot of pseudogenes (67). For 50 ncRNAs such as CAESAR (RF00172), G-CSF_SLIDE (RF00183), NRON (RF00636), TUSC7 (RF01879), Xist_exon4 (RF01881), LINC00901 (RF01884), and Hammerhead_HH9 and _HH10 (RF02275, RF02277), we found exactly one copy in each investigated bat genome assembly (Supplementary Table S8). Again, we also observed ncRNA families that are lost for some species or entire families, for example the ribozyme CoTC (RF00621) that seems to be absent from all *Vespertilionidae* members and *M. natalensis*.

Mitochondrial annotation

For each investigated bat species except *E. spelaea*, *R. sinicus*, *M. lyra*, *E.fuscus* and *M. natalensis* (where no mitochondrial contigs could be identified; Table 3) mitochondrial protein-coding genes and ncRNAs were annotated (see Methods). In total, 37 mitochondrial genes comprising 22 tRNAs, two rRNAs (12S and 16S) and 13 protein-coding genes were detected for each bat species as known for other metazoans (68) (Supplementary Table S10). The mitochondrial genome lengths range from 16 343 nt in *E. helvum* to 17 783 nt in *M. davidii*. For the five bat species, where the mitochondrial genome could be identified as a part of the NCBI genome assembly, we appended the mtDNA annotation to the final annotation of ncRNAs.

Updated annotations provide insights into novel differentially expressed ncRNAs

Exemplarily, we investigated known and novel differentially expressed (DE) ncRNAs found in the genome of *M. lucifugus* in more detail. To this end, we used the RNA-Seq data sets *Field-2015*, *Field-2018*, *Hölzer-2019* and *Weber-2019* (Table 2) as a basis to identify DE ncRNAs that were newly discovered in this study and were not part of the current NCBI or Ensembl genome annotations for this bat species. More detailed DE results can be found in Supplementary File S2.

We filtered for novel *M. lucifugus* DE ncRNAs by (i) an absolute $\log_2$ fold change (fc) > 1, (ii) an adjusted p-value < 0.05, and (iii) a TPM > 10. We further manually investigated the expression patterns with the IGV (69) and discarded DE ncRNAs overlapping with the current NCBI (Myoluc2.0 annotation release 102) or Ensembl (Myoluc2.0.96) annotations.

Based on the small RNA-Seq comparison of mock and virus-infected (Clone 13) samples 24 h post infection (*Weber-2019*), we found several miRNAs (Rfam- and mirDeep2-based) and snoRNAs to be differentially expressed (Figure 4 and details in Supplementary Files S2.13). In general, replicates of virus-infected and IFN-treated samples 24 h post infection tend to cluster together only based on the expression profiles of small ncRNAs (mainly miRNAs) (Figure 4A and B). Most differences can be observed between the 24 h virus-infected and all other samples, which seem to show no clearly distinguishable expression pattern. Interestingly, at 6 h post infection, we see replicates clustering together regardless of their treatment (mock, IFN, Clone 13). Thus, after only 6 h, few miRNAs are differentially expressed and therefore the samples of each replicate (mock, IFN, Clone 13) cluster together, because they have the same passaging history but the passaging history in between the replicates differ (26). After 24 h, more and more miRNAs are significantly differentially expressed and the samples can be better distinguished based on their treatment (Figure 4A and B). We observed that, in general, miRNAs tend to be down-regulated (Figure 4B; upper half), while snoRNAs tend to be up-regulated (lower half) after 24 h of Clone 13 infection compared to mock. For example, we found a novel miRNA (MLUGD00000002094 in our annotation; predicted by mirDeep2; Supplementary Table S7) located in an intron of the protein-coding gene *SEMA3G*, significantly down-regulated ($\log_2$ fc $= -2.56$) during Clone 13 infection (Figure 4C and D). Based on Rfam alignments we further found a histone 3'-UTR stem-loop (RF00032), an RNA element involved in nucleocytoplasmic transport of the histone mRNAs, significantly down-regulated during infection.

For the same comparison of mock and virus-infected samples at 24 h, the rRNA-depleted data set (*Hölzer-2019*) revealed several DE lncRNAs. For example, we found a lncRNA potentially transcribed in an intron of *MX1* (MLUGL00000039178) up-regulated ($\log_2$ fc $= 3.36$) during infection. Another lncRNA (MLUGL00000087396), we found potentially transcribed as a part of two exons of the *PLAT* protein-coding gene and down-regulated ($\log_2$ fc $= -1.67$) during viral infection (see details in Supplementary Files S2.11).

Interestingly, based on the *Field-2015* RNA-Seq data, we found two internal ribosomal entry site (IRES) in the genes *VEGFA* and *ODC* with Rfam IDs RF00461 and RF02535, respectively, to be 2-fold up-regulated during *P. destructans* infection.

DISCUSSION

In this study, we comprehensively annotated ncRNAs in 16 readily available bat genomes obtained from the NCBI database (Table 1). We provide novel annotations in the common GTF format, following a hierarchical structure of *gene*, *transcript* and *exon* features to allow direct integration of our annotations into already available ones (Supplementary Tables S3–S10). Finally, we provide for each bat genome assembly an extended annotation file merged with the protein- and non-coding gene annotations that were already available by the NCBI database (Supplementary Files

S1; leaving out potential lncRNAs that can be downloaded separately, see Supplementary Table S9). We used six RNA-Seq data sets derived from the transcriptomic sequencing of four bat species (Table 2) to calculate normalized expression values for our newly annotated ncRNAs and exemplarily show significantly differential expressed ncRNAs (Figure 4), which were never before described on such a large scale for any bat species.

In addition to the evolutionarily explainable differences in the pure existence and the amount of annotated ncRNAs in bats, we have observed that the assembly quality can also influence the annotation results. While the effects on the annotation of short ncRNAs seem to be small (with some exceptions), the number of identified lncRNAs and lncRNA hot spots increases with increasing assembly quality (e.g. with a higher N50; see Supplementary Figure S1). However, this observation may be true for our data and analyses, but it also depends strongly on the annotation method used.

Recently, the Bat1K project (http://bat1k.ucd.ie/) was announced as a global effort to sequence, assemble and annotate high-quality genomes of all living bat species (2). We aim to extend our annotation of ncRNAs regularly and whenever new bat genomes become publicly available.

In mid January 2019, 16 new low-coverage bat genomes of nine families were submitted by the Broad Institute to the NCBI genome database. Unfortunately, our time-consuming and computationally extensive analyses were already completed at this time point. We want to further automate our ncRNA annotation workflow, to easily include these and any new bat (or other mammalian) genomes that will be sequenced and assembled in the future.

Our current identification of ncRNAs in bat species will be usable as a resource (Electronic Supplement) for deeper studying of bat evolution, ncRNAs repertoire, gene expression and regulation, ecology, and important host–virus interactions.

DATA AVAILABILITY

Detailed information about the bat genomes used in this study, their assembly quality and all ncRNA candidates (in FASTA, STK and GTF format) can be found in the Electronic Supplement (rna.uni-jena.de/supplements/bats). The final extended annotations for each investigated bat species can be found in Supplementary Files S1 and the lncRNA annotations in Supplementary Table S9. To allow full reproducibility of our study, all final and intermediate data files (such as used genome files and mapping files in BAM format) were uploaded to the Open Science Framework under accession doi.org/10.17605/OSF.IO/4CMDN. Python scripts used to filter and merge our annotations were deposited at GitHub (github.com/rnajena/bats_ncrna). The virus-infected and IFN-stimulated small RNA-Seq data of the *M. daubentonii* kidney cell line was uploaded to GEO (GSE132336).

ACKNOWLEDGEMENTS

We thank Ivonne Görlich and Marco Groth from the Core Facility DNA sequencing of the Leibniz Institute on Aging – Fritz Lipmann Institute in Jena for their help with short-read Illumina sequencing. Martin Hölzer appreciates the support of the Joachim Herz Foundation by the add-on fellowship for interdisciplinary life science. Kevin Lamkiewicz thanks the Carl Zeiss Foundation for the financial support within the scope of the program line ”Breakthroughs”. The authors gratefully acknowledge the help of the developer of the Gorap pipeline, Konstantin Riege.

FUNDING

Deutsche Forschungsgemeinschaft SFB 1021 (B06), SPP 1596 [We 2616/7-1, We 2616/7-2, MA 5082/7-1, MA 5082/9-1]; LOEWE-Schwerpunkt ‘Medical RNomics’ of the Land Hessen and the RAPID consortium of the Bundesministerium für Bildung und Forschung (BMBF) [01KI1723E]; Friedrich Schiller University Jena, Collaborative Research Centre AquaDiva [CRC 1076 AquaDiva]; University of Leipzig, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig [iDIV FZT 118]; Carl Zeiss Foundation, project ‘Eine virtuelle Werkstatt für die Digitalisierung in den Wissenschaften’ [Durchbrüche].
Conflict of interest statement. None declared.

REFERENCES

1. Calisher,C.H., Childs,J.E., Field,H.E., Holmes,K.V. and Schountz,T. (2006) Bats: important reservoir hosts of emerging viruses. *Clin. Microbiol. Rev.*, **19**, 531–545.
2. Teeling,E.C., Vernes,S.C., Dávalos,L.M., Ray,D.A., Gilbert,M.T.P., Myers,E. and Consortium,B. (2018) Bat biology, genomes, and the Bat1K Project: to generate chromosome-level genomes for all living bat species. *Annu. Rev. Anim. Biosci.*, **6**, 23–46.
3. Simmons,N.B. (2005) An Eocene big bang for bats. *Science*, **307**, 527–528.
4. Teeling,E.C., Jones,G. and Rossiter,S.J. (2016) Phylogeny, genes, and hearing: implications for the evolution of echolocation in bats. In: *Bat Bioacoustics*. Springer, pp. 25–54.
5. Gardner,A., Wilson,D. and Reeder,D. (2005) Mammal species of the world. A taxonomic and geographic reference. *Mammal Species World*, **12**, 312–529.
6. Teeling,E.C., Springer,M.S., Madsen,O., Bates,P., O'brien,S.J. and Murphy,W.J. (2005) A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science*, **307**, 580–584.
7. Prat,Y., Azoulay,L., Dor,R. and Yovel,Y. (2017) Crowd vocal learning induces vocal dialects in bats: Playback of conspecifics shapes fundamental frequency usage by pups. *PLoS Biol.*, **15**, e2002556.
8. Huang,Z., Jebb,D. and Teeling,E.C. (2016) Blood miRNomes and transcriptomes reveal novel longevity mechanisms in the long-lived bat, Myotis myotis. *BMC Genom.*, **17**, 906.
9. Foley,N.M., Hughes,G.M., Huang,Z., Clarke,M., Jebb,D., Whelan,C.V., Petit,E.J., Touzalin,F., Farcy,O., Jones,G. *et al.* (2018) Growing old, yet staying young: The role of telomeres in bats' exceptional longevity. *Sci. Adv.*, **4**, eaao0926.
10. Huang,Z., Whelan,C.V., Foley,N.M., Jebb,D., Touzalin,F., Petit,E.J., Puechmaille,S.J. and Teeling,E.C. (2019) Longitudinal comparative transcriptomics reveals unique mechanisms underlying extended healthspan in bats. *Nat. Ecol. Evol.*, **3**, 1110–1120.
11. Smith,I. and Wang,L.-F. (2013) Bats and their virome: an important source of emerging viruses capable of infecting humans. *Curr. Opin. Virol.*, **3**, 84–91.
12. Wang,L.-F., Walker,P.J. and Poon,L.L. (2011) Mass extinctions, biodiversity and mitochondrial function: are bats ‘special’ as reservoirs for emerging viruses? *Curr. Opin. Virol.*, **1**, 649–657.
13. Brook,C.E. and Dobson,A.P. (2015) Bats as ‘special’ reservoirs for emerging zoonotic pathogens. *Trends Microbiol.*, **23**, 172–180.
14. Anthony,S.J., Johnson,C.K., Greig,D.J., Kramer,S., Che,X., Wells,H., Hicks,A.L., Joly,D.O., Wolfe,N.D., Daszak,P. *et al.* (2017) Global patterns in coronavirus diversity. *Virus Evol.*, **3**, vex012.
15. O'shea,T.J., Cryan,P.M., Cunningham,A.A., Fooks,A.R., Hayman,D.T., Luis,A.D., Peel,A.J., Plowright,R.K. and Wood,J.L.

(2014) Bat flight and zoonotic viruses. *Emerg. Infect. Dis.*, **20**, 741–745.

16. Hölzer,M., Krähling,V., Amman,F., Barth,E., Bernhart,S.H., Carmelo,V.A.O., Collatz,M., Doose,G., Eggenhofer,F., Ewald,J. *et al.* (2016) Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells. *Sci. Rep.*, **6**, 34589.
17. Papenfuss,A.T., Baker,M.L., Feng,Z.-P., Tachedjian,M., Crameri,G., Cowled,C., Ng,J., Janardhana,V., Field,H.E. and Wang,L.-F. (2012) The immune gene repertoire of an important viral reservoir, the Australian black flying fox. *BMC Genom.*, **13**, 261.
18. Mattick,J. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep*, **2**, 986–991.
19. Isakov,O., Ronen,R., Kovarsky,J., Gabay,A., Gan,I., Modai,S. and Shomron,N. (2012) Novel insight into the non-coding repertoire through deep sequencing analysis. *Nucleic Acids Res*, **40**, e86.
20. Stein,L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, **2**, 493–503.
21. Ekblom,R. and Galindo,J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.
22. Salzberg,S.L. (2019) Next-generation genome annotation: we still struggle to get it right. *Genome Biol.*, **20**, 92.
23. Haynes,W.A., Tomczak,A. and Khatri,P. (2018) Gene annotation bias impedes biomedical research. *Sci. Rep.*, **8**, 1362.
24. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
25. Eckalbar,W.L., Schlebusch,S.A., Mason,M.K., Gill,Z., Parker,A.V., Booker,B.M., Nishizaki,S., Muswamba-Nday,C., Terhune,E., Nevonen,K.A. *et al.* (2016) Transcriptomic and epigenomic characterization of the developing bat wing. *Nat. Genet.*, **48**, 528.
26. Hölzer,M., Schoen,A., Wulle,J., Müller,M.A., Drosten,C., Marz,M. and Weber,F. (2019) Virus- and interferon alpha-induced transcriptomes of cells from the microbat *Myotis daubentonii*. *iScience*, **19**, 647–661.
27. Field,K.A., Johnson,J.S., Lilley,T.M., Reeder,S.M., Rogers,E.J., Behr,M.J. and Reeder,D.M. (2015) The white-nose syndrome transcriptome: activation of anti-fungal host responses in wing tissue of hibernating little brown myotis. *PLOS Pathog*, **11**, e1005168.
28. Field,K.A., Sewall,B.J., Prokkola,J.M., Turner,G.G., Gagnon,M.F., Lilley,T.M., Paul White,J., Johnson,J.S., Hauer,C.L. and Reeder,D.M. (2018) Effect of torpor on host transcriptomic responses to a fungal pathogen in hibernating bats. *Mol Ecol*, **27**, 3727–3743.
29. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina Sequence Data. *Bioinformatics*, **30**, 2114–2120.
30. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
31. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
32. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10.
33. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
34. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
35. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
36. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
37. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2017) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
38. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
39. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
40. Kalvari,I., Nawrocki,E.P., Argasinska,J., Quinones-Olvera,N., Finn,R.D., Bateman,A. and Petrov,A.I. (2018) Non-coding RNA analysis using the Rfam database. *Curr. Protoc. Bioinformatics*, p. e51.
41. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
42. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
43. Griffiths-Jones,S. (2005) RALEE—RNA Alignment editor in Emacs. *Bioinformatics*, **21**, 257–259.
44. Lagesen,K., Hallin,P., Rødland,E.A., Staerfeldt,H.-H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
45. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
46. Kehr,S., Bartschat,S., Tafer,H., Stadler,P.F. and Hertel,J. (2013) Matching of Soulmates: coevolution of snoRNAs and their targets. *Mol. Biol. Evol.*, **31**, 455–467.
47. Cowled,C., Stewart,C.R., Likic,V.A., Friedländer,M.R., Tachedjian,M., Jenkins,K.A., Tizard,M.L., Cottee,P., Marsh,G.A., Zhou,P. *et al.* (2014) Characterisation of novel microRNAs in the Black flying fox (*Pteropus alecto*) by deep sequencing. *BMC Genom.*, **15**, 682.
48. Volders,P.-J., Anckaert,J., Verheggen,K., Nuytens,J., Martens,L., Mestdagh,P. and Vandesompele,J. (2019) LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D135–D139.
49. Bernt,M., Donath,A., Jühling,F., Externbrink,F., Florentz,C., Fritzsch,G., Pütz,J., Middendorf,M. and Stadler,P.F. (2013) MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.*, **69**, 313–319.
50. Khan,A.R., Pervez,M.T., Babar,M.E., Naveed,N. and Shoaib,M. (2018) A comprehensive study of de novo genome assemblers: current challenges and future prospective. *Evol. Bioinform.*, **14**, doi:10.1177/1176934318758650.
51. Parker,J., Tsagkogeorga,G., Cotton,J.A., Liu,Y., Provero,P., Stupka,E. and Rossiter,S.J. (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**, 228–231.
52. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
53. Zhang,G., Cowled,C., Shi,Z., Huang,Z., Bishop-Lilly,K.A., Fang,X., Wynne,J.W., Xiong,Z., Baker,M.L., Zhao,W. *et al.* (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science*, **339**, 456–460.
54. Dong,D., Lei,M., Hua,P., Pan,Y.-H., Mu,S., Zheng,G., Pang,E., Lin,K. and Zhang,S. (2016) The genomes of two bat species with long constant frequency echolocation calls. *Mol. Biol. Evol.*, **34**, 20–34.
55. Mendoza,M. L.Z., Xiong,Z., Escalera-Zamudio,M., Runge,A.K., Thézé,J., Streicker,D., Frank,H.K., Loza-Rubio,E., Liu,S., Ryder,O.A. *et al.* (2018) Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. *Nat. Ecol. Evol.*, **2**, 659–668.
56. Seim,I., Fang,X., Xiong,Z., Lobanov,A.V., Huang,Z., Ma,S., Feng,Y., Turanov,A.A., Zhu,Y., Lenz,T.L. *et al.* (2013) Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nat. Commun.*, **4**, 2212.
57. Wen,M., Ng,J.H., Zhu,F., Chionh,Y.T., Chia,W.N., Mendenhall,I.H., Lee,B. P.-H., Irving,A.T. and Wang,L.-F. (2018) Exploring the genome and transcriptome of the cave nectar bat Eonycteris spelaea with PacBio long-read sequencing. *GigaScience*, **7**, giy116.
58. Pavlovich,S.S., Lovett,S.P., Koroleva,G., Guito,J.C., Arnold,C.E., Nagle,E.R., Kulcsar,K., Lee,A., Thibaud-Nissen,F., Hume,A.J. *et al.* (2018) The Egyptian rousette genome reveals unexpected features of bat antiviral immunity. *Cell*, **173**, 1098–1110.
59. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2015) Reference sequence (RefSeq) database at [NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
60. Shaw,T.I., Srivastava,A., Chou,W.-C., Liu,L., Hawkinson,A., Glenn,T.C., Adams,R. and Schountz,T. (2012) Transcriptome

sequencing and annotation for the Jamaican fruit bat (*Artibeus jamaicensis*). *PLoS One*, **7**, e48472.

61. Yuan,L., Geiser,F., Lin,B., Sun,H., Chen,J. and Zhang,S. (2015) Down but not out: the role of microRNAs in hibernating bats. *PLoS One*, **10**, e0135064.
62. Schattner,P., Barberan-Soler,S. and Lowe,T.M. (2006) A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, **12**, 15–25.
63. McMahon,M., Contreras,A. and Ruggero,D. (2015) Small RNAs with big implications: new insights into H/ACA snoRNA function and their role in human disease. *Wiley Interdisciplin. Rev.*, **6**, 173–189.
64. Kung,J.T.Y., Colognori,D. and Lee,J.T. (2013) Long noncoding RNAs: past, present, and future. *Genetics*, **193**, 651–669.
65. Nam,J.-W., Choi,S.-W. and You,B.-H. (2016) Incredible RNA: dual functions of coding and noncoding. *Mol. Cell*, **39**, 367–374.
66. Kumari,P. and Sampath,K. (2015) cncRNAs: Bi-functional RNAs with protein coding and non-coding functions. *Semin. Cell Dev. Biol.*, **47-48**, 40–51.
67. Marz,M., Kirsten,T. and Stadler,P.F. (2008) Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, **67**, 594–607.
68. Boore,J.L. (1999) Animal mitochondrial genomes. *Nucleic Acids Res.*, **27**, 1767–1780.
69. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinf.*, **14**, 178–192.
70. Mikheenko,A., Valin,G., Prjibelski,A., Saveliev,V. and Gurevich,A. (2016) Icarus: visualizer for *de novo* assembly evaluation. *Bioinformatics*, **32**, 3321–3323.