Annotation and differential expression analysis of non-coding RNAs in 16 freely accessible bat genomes

Marie Lataretu, Friedrich-Schiller-Universität Jena

14th February, 2019

Bats are cool!

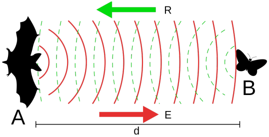
Features

- The only flying mammals

Bats are cool!

Features

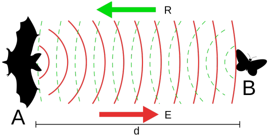
- The only flying mammals
- Laryngeal echolocation



Bats are cool!

Features

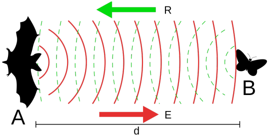
- The only flying mammals
- Laryngeal echolocation
- Vocal learning



Bats are cool!

Features

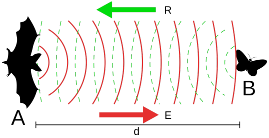
- The only flying mammals
- Laryngeal echolocation
- Vocal learning
- Account for $\sim$ 20 % of all mammal species

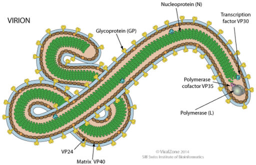


Bats are cool!

Features

- The only flying mammals
- Laryngeal echolocation
- Vocal learning
- Account for ~ 20 % of all mammal species
- Immunity against various pathogenic viruses

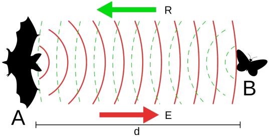


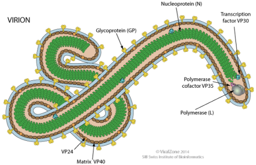


Bats are cool!

Features

- The only flying mammals
- Laryngeal echolocation
- Vocal learning
- Account for ~ 20 % of all mammal species
- Immunity against various pathogenic viruses
- Show homosexual behavior[1]

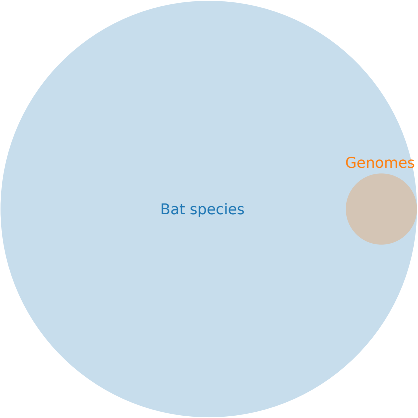




[1] B. Bagemihl. Biological Exuberance: Animal Homosexuality and Natural Diversity. 1999.

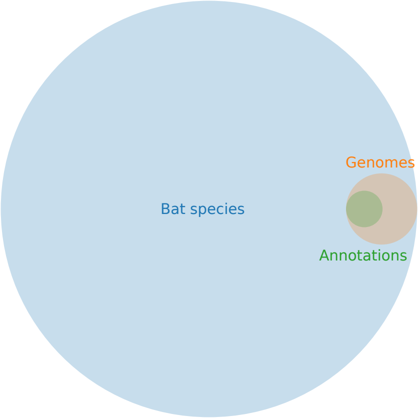
Freely available genomes and annotations (today)

- Genomes: 32 of more than 1,300 species

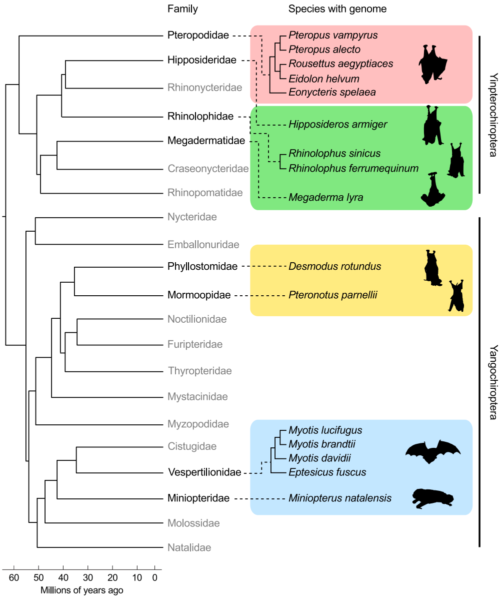


Freely available genomes and annotations (today)

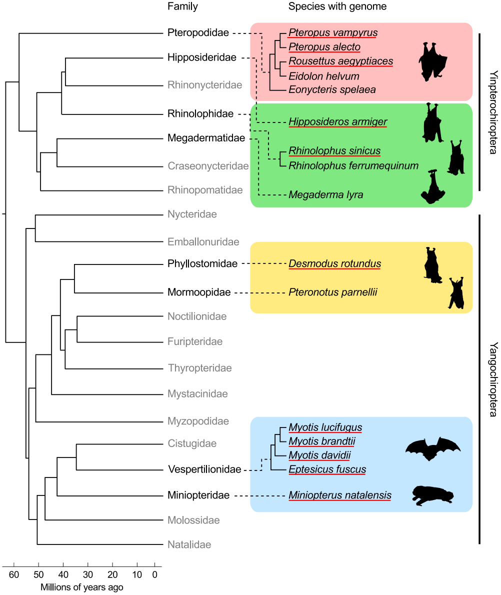
- Genomes: 32 of more than 1,300 species
- Annotations: 11 of 32 species



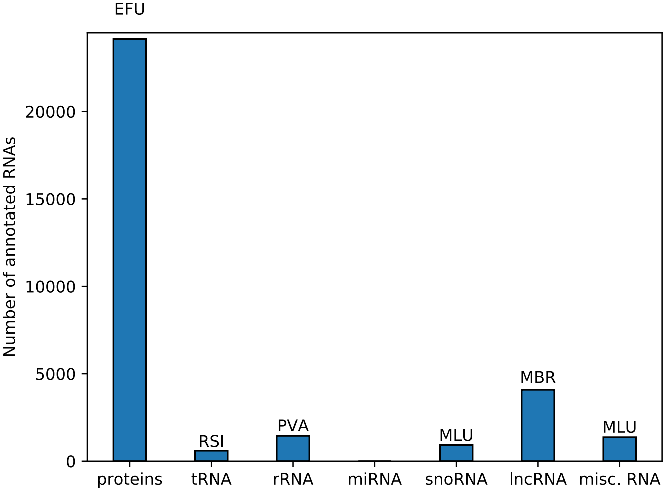
Freely available genomes (before 15 January 2019)



Freely available annotations (before 15 January 2019)



Freely available annotations (before 15 January 2019)



Maximal number of annotated RNAs for each RNA class.

Hackaton

Estonteco
FULL FORCE
HACKEN
Estonteco
HACKEN
HACKEN
HACKEN

Hackaton

1. Annotation of non-coding RNAs in 16 bats

Hackaton

1. Annotation of non-coding RNAs in 16 bats
2. Differential expression analysis of non-coding RNAs

Annotation of ncRNA in 16 bats

Coordinator

Martin

Annotation of ncRNA in 16 bats

Coordinator

Martin

ncRNA classes

- tRNAs
- snoRNAs
- miRNAs
- lncRNAs
- Mitochondrial annotation
- And others (e.g. snRNAs)

Annotation of ncRNA in 16 bats

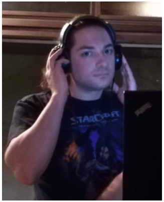
rRNA

1. `RNAmmer` (v1.2)[2]
   - Hidden markov models

→ 5.8S, 18S and 28S rRNA

[2] K. Lagesen et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. 2007.

Annotation of ncRNA in 16 bats



tRNA

1. `tRNAscan-SE` [3]
   - Default parameters
2. Remove ‘Undet’ or ‘Pseudo’ types

[3] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. 1997.

Annotation of ncRNA in 16 bats

snoRNA, miRNA and others

1. `Gorap`[4] (uses `Infernal`) with alignments from the `Rfam`[5] data base

[4] github.com/koriege/gorap

[5] I. Kalvari et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. 2017.

Annotation of ncRNA in 16 bats

snoRNA, miRNA and others

1. Gorap[4] (uses Infernal) with alignments from the Rfam[5] data base
2. For snoRNAs:
   - Classification of C/D box and H/ACA box

[4] github.com/koriege/gorap

[5] I. Kalvari et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. 2017.

Annotation of ncRNA in 16 bats

miRNA

- `miRDeep2` (v2.0.0.8)[6]
  - Input:
    - Combined smallRNA-Seq data set[7]
    - Mapped to each individual bat assembly

[6] M. R. Friedländer et al. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. 2012.

[7] Unpublished data, provided by Friedemann Weber, Justus-Liebig-Universität Giessen

Annotation of ncRNA in 16 bats

lncRNA

Data

- LNCipedia (v5.2)[8] data base
  - High confidence set:
    - 107,039 transcript of potential human lncRNAs

[8] P.-J. Volders et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. 2019.

[9] S. F. Altschul et al. Basic local alignment search tool. 1990.

Annotation of ncRNA in 16 bats

lncRNA

Data

- `LNCipedia` (v5.2)[8] data base
  - High confidence set:
    - 107,039 transcript of potential human lncRNAs

Tool

1. `BLASTn`[9] (v2.7.1+, $1e^{-10}$)

[8] P.-J. Volders et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. 2019.

[9] S. F. Altschul et al. Basic local alignment search tool. 1990.

Annotation of ncRNA in 16 bats

lncRNA

Data

- `LNCipedia` (v5.2)[8] data base
  - High confidence set:
    - 107,039 transcript of potential human lncRNAs

Tool

1. `BLASTn`[9] (v2.7.1+, 1e^{-10})
2. Filter and re-structure the result

   → Gene - transcript - exon structure

   → Indroduce *lncRNA hot spots*

[8] P.-J. Volders et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. 2019.

[9] S. F. Altschul et al. Basic local alignment search tool. 1990.

Annotation of ncRNA in 16 bats

Mitochondrial annotation

Data

- 10 `NCBI` mitogenomes
- 1 blasted mitogenome
  - Rearrange the mitogenome

[10] M. Bernt et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. 2013.

Annotation of ncRNA in 16 bats

Mitochondrial annotation

Data

- 10 NCBI mitogenomes
- 1 blasted mitogenome
  - Rearrange the mitogenome

Tool

- MITOS2 [10]
  → Protein coding and non-coding RNA annotation

[10] M. Bernt et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. 2013.

Annotation of ncRNA in 16 bats

Finalization

- Check `gtf` format

Annotation of ncRNA in 16 bats

Finalization

- Check `gtf` format
- Merge all annotations for each bat
  - Check for overlaps:
    1. Within the new annotations
    2. In the existing `NCBI` annotations

Annotation of ncRNA in 16 bats

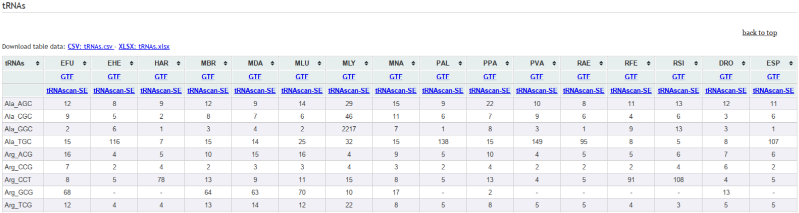
Finalization

- Check `gtf` format
- Merge all annotations for each bat
  - Check for overlaps:
    1. Within the new annotations
    2. In the existing `NCBI` annotations
- Produce nice `html` tables for each annotation
  - Automated `csv` and `xlsx` generation

Annotation of ncRNA in 16 bats

Finalization

- Check `gtf` format
- Merge all annotations for each bat
  - Check for overlaps:
    1. Within the new annotations
    2. In the existing `NCBI` annotations
- Produce nice `html` tables for each annotation
  - Automated `csv` and `xlsx` generation



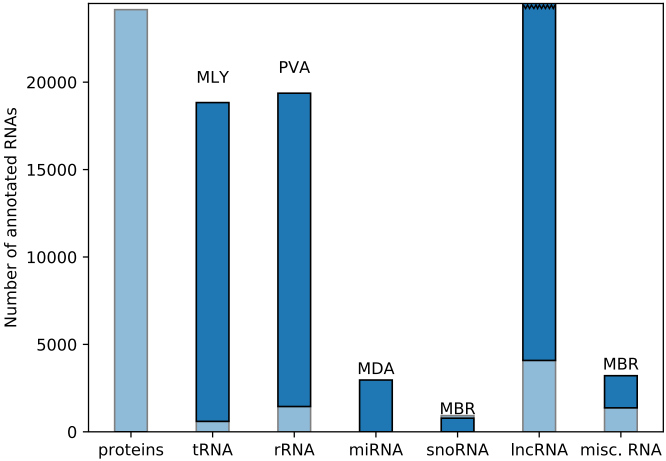
tRNAs

back to top

Download table data: CSV: tRNAs.csv - XLSX: tRNAs.xlsx

tRNAs	EFU	EHE	HAR	MBR	MDA	MLU	MLY	MNA	PAL	PPA	PVA	RAE	RFE	RSI	DRO	ESP
	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF	GTF
	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE	tRNAscan-SE
Ala_AGC	12	8	9	12	9	14	29	15	9	22	10	8	11	13	12	11
Ala_CGC	9	5	2	8	7	6	46	11	6	7	9	6	4	6	3	6
Ala_GGC	2	6	1	3	4	2	2217	7	1	8	3	1	9	13	3	1
Ala_TGC	15	116	7	15	14	25	32	15	138	15	149	95	8	5	8	107
Arg_ACG	16	4	5	10	15	16	4	9	5	10	4	5	5	6	7	6
Arg_CCG	7	2	4	2	3	3	4	3	2	4	2	2	2	4	6	2
Arg_CCT	8	5	78	13	9	11	15	8	5	13	4	5	91	108	4	5
Arg_GCG	68	-	-	64	63	70	10	17	-	2	-	-	-	-	13	-
Arg_TCG	12	4	4	13	14	12	22	8	5	8	5	5	4	3	5	5

Results



Maximal number of newly annotated RNAs for each RNA class. Newly annotated lncRNAs: 286805

Results

- Final annotation for each bat in `gft` format
- Annotations for each ncRNA class and bat
  → Compatible and useable annotations

Hackaton

1. Annotation of non-coding RNAs in 16 bats
2. **Differential expression analysis of non-coding RNAs**

Differential expression analysis of non-coding RNAs

Data

- 6 RNA-Seq data sets
  - 98 samples in total
  - From 4 different bat species

Differential expression analysis of non-coding RNAs

Pipeline

- Preprocessing with `Trimmomatic` (v0.36) [11]

[11] A. M. Bolger et al. Trimmomatic: A flexible trimmer for Illumina sequence data. 2014.

[12] D. Kim et al. HISAT: a fast spliced aligner with low memory requirements. 2015.

Differential expression analysis of non-coding RNAs

Pipeline

- Preprocessing with `Trimmomatic` (v0.36) [11]
- Mapping with `HISAT` (v2.1.0) [12]
  - Each sample individually
    → 1568 mappings in total

[11] A. M. Bolger et al. Trimmomatic: A flexible trimmer for Illumina sequence data. 2014.

[12] D. Kim et al. HISAT: a fast spliced aligner with low memory requirements. 2015.

[13] Y. Liao et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. 2014.

Differential expression analysis of non-coding RNAs

Pipeline

- Preprocessing with `Trimmomatic` (v0.36) [11]
- Mapping with `HISAT` (v2.1.0) [12]
  - Each sample individually
    → 1568 mappings in total
- Counting with `featureCounts` (v1.6.3)[13]
  - Only unique mapped reads

[11] A. M. Bolger et al. Trimmomatic: A flexible trimmer for Illumina sequence data. 2014.

[12] D. Kim et al. HISAT: a fast spliced aligner with low memory requirements. 2015.

[13] Y. Liao et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. 2014.

Differential expression analysis of non-coding RNAs

Analysis

- Differential gene expression analyses with `DESeq2` [14]
  - `DESeq2` normalization
    → Pairwise comparisons
    → Significantly[15] differentially expressed ncRNAs

[14] M. I. Love et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 2014.

[15] Adjusted p-value < 0.05; absolute log 2 fold change > 2

Differential expression analysis of non-coding RNAs

Analysis

- Differential gene expression analyses with `DESeq2` [14]
    - `DESeq2` normalization
      → Pairwise comparisons
      → Significantly[15] differentially expressed ncRNAs
- TPM (transcripts per million) for each ncRNA in each sample
  → Normalized expression level of each ncRNA

[14] M. I. Love et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 2014.

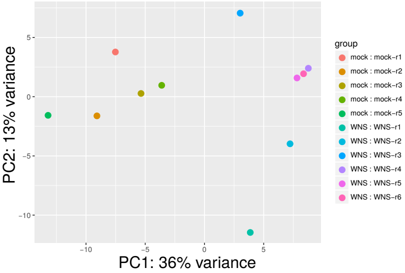
[15] Adjusted p-value < 0.05; absolute log 2 fold change > 2

Preliminary results

- RNA-Seq data set: *Field-2015*[16]
  - 5 mock samples
  - 6 infected (white-nose syndrome, WNS) samples

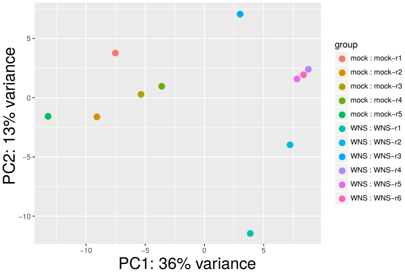
[16] K. A. Field et al. The white-nose syndrome transcriptome: activation of anti-fungal host responses in wing tissue of hibernating little brown myotis. 2015.

Preliminary results

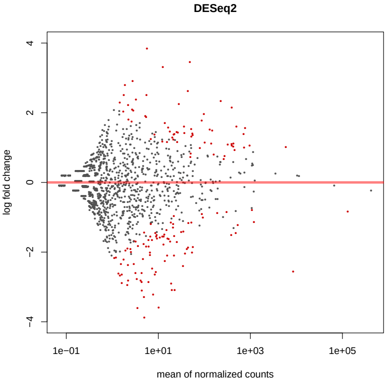


PCA on ncRNAs.

Preliminary results

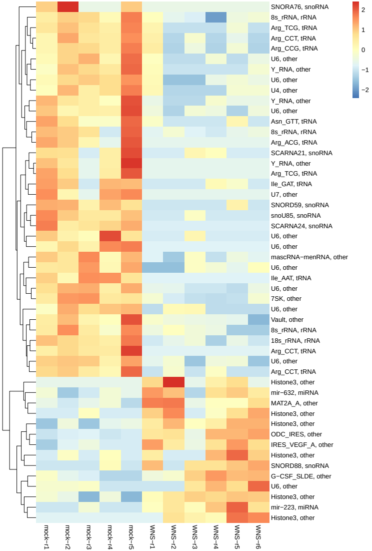


PCA on ncRNAs.



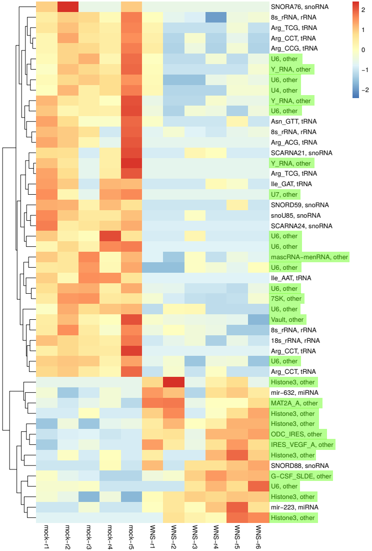
MA plot.

Preliminary results



Expression levels of significantly differentially expressed genes.

Preliminary results



Expression levels of significantly differentially expressed genes.

What is next?

- Analyze the other RNA-Seq data sets
- Make the annotations and results available

[17] https://bat1k.ucd.ie/

What is next?

- Analyze the other RNA-Seq data sets
- Make the annotations and results available

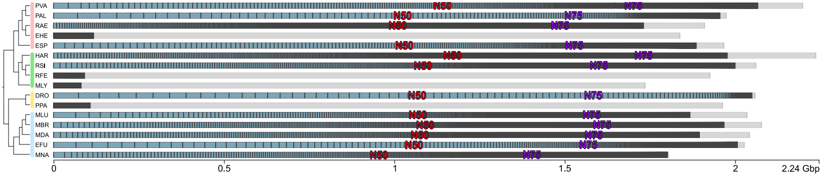
- Hack the 16 new `NCBI` assemblies
- Bat1K project[17]: sequence the genomes of all living bat species

[17] https://bat1k.ucd.ie/

Thanks to

- Manja Marz
- Martin Hölzer
- Nelly Fernanda Mostajo Berrospi
- RNA Bioinformatics & High-Throughput Analysis Jena

Genome quality



Icarus plot of the 16 investigated bat species: assembly lengths, N50 and N75 values.