UNIVERSITÄT BONN

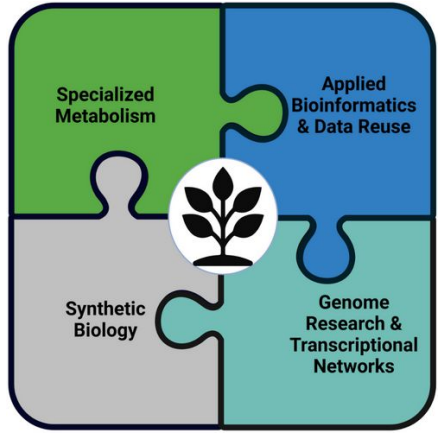
PuckerLab

de.NBI course 2: From Gene Models to Biological Insights - Functional Annotation

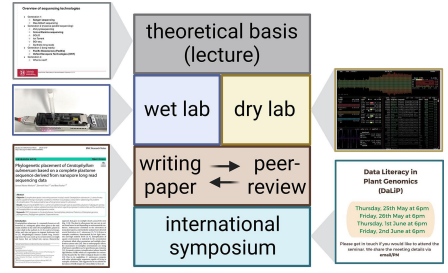
@PuckerLab

@puckerlab

- Flavonoids: Anthocyanins, Flavonols, Flavones, Proanthocyanidins
- Withanolides
- Carotenoids
- …

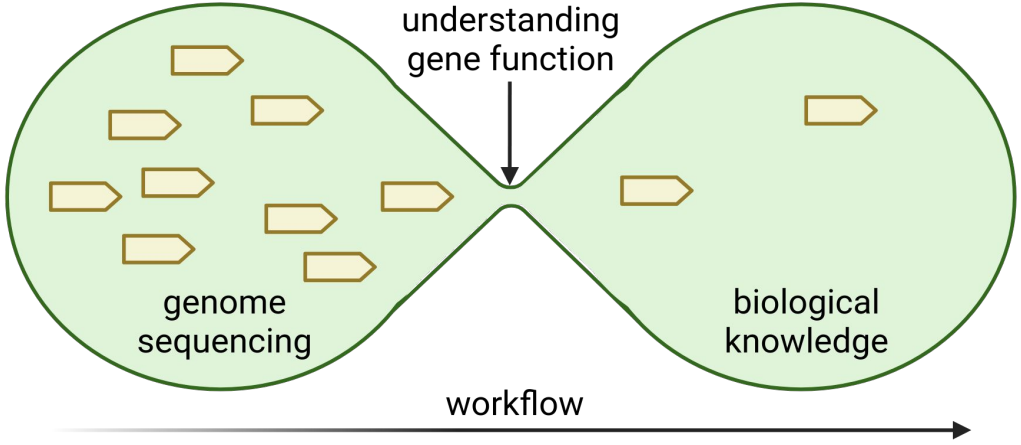


Tools: KIPES, NAVIP, MGSE, MYB/bHLH_annotator, DupyliCate…



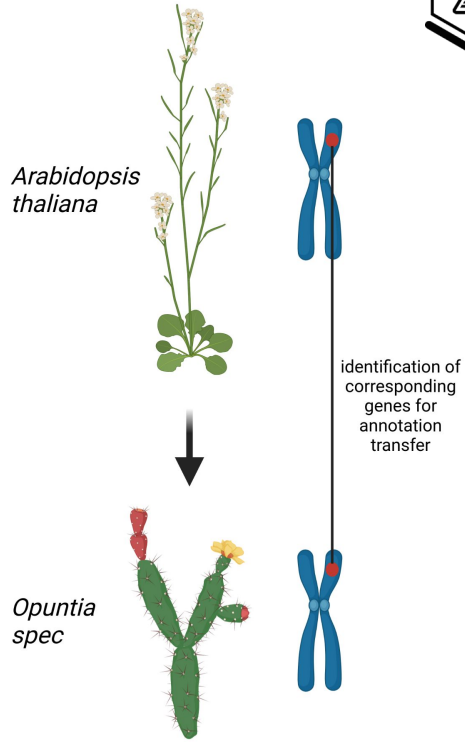
Wolff et al., 2023: 10.1515/jib-2023-0033
Dassow et al., 2025: 10.1101/2025.01.31.635922

Course motivation



From L to R: Shakunthala, Samuel, Nancy, Julie, and Boas

Picture credits: Jakob Horz



Pucker, 2024: 10.20944/preprints202402.0645.v1

Availability of Slides

All materials will be shared with participants at the end of this course. Background information is available here:

https://www.preprints.org/manuscript/202508.1176

Slides →https://github.com/bpucker/teaching/tree/master/deNBI2025

UNIVERSITÄT BONN

Start of Day 1

Today you will learn:

- Introduction to Virtual Machine
- Access to Virtual Machine
- Running BUSCO
- Transposable element annotation

Introduction to Virtual Machines (VM)

- high performance computing => server farm

UNIVERSITÄT BONN

Introduction to Virtual Machines (VM)

- high performance computing => server farm
- de.NBI offers VMs (VM = virtual machine)
  - running Linux
  - access through SSH

- de.NBI connected LifeScience AAI Account needed

Introduction to Virtual Machines (VM)

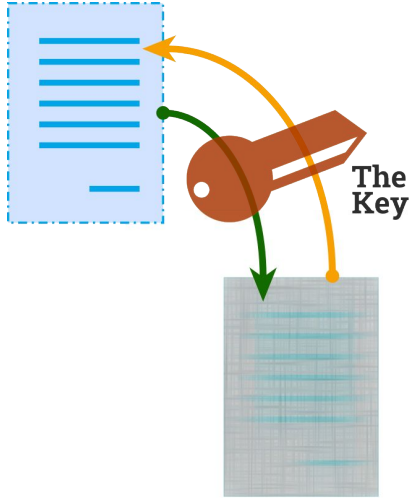
- high performance computing => server farm
- de.NBI offers VMs (VM = virtual machine)
  - running Linux
  - access through SSH

Who is new?

- de.NBI connected LifeScience AAI Account needed

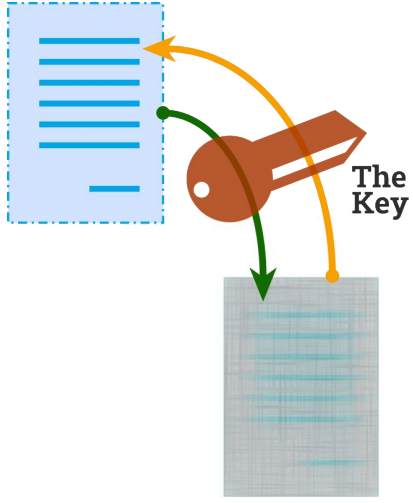
- Tasks:
  1. Register: https://cloud.denbi.de/wiki/registration/
  2. Apply for the course de.NBI project: https://signup.aai.lifescience-ri.eu/fed/registrar/?vo=denbi&group=pbbcourse20251

SSH keys - asymmetric encryption

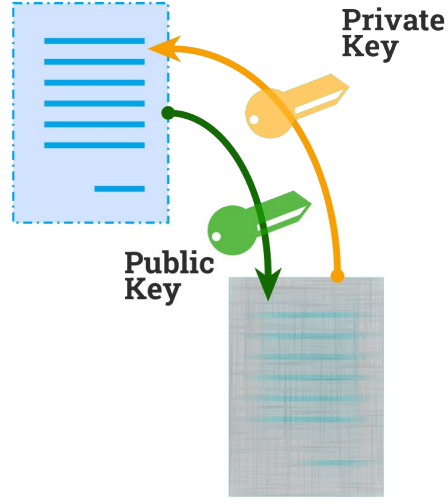


Symmetric encryption

SSH keys - asymmetric encryption

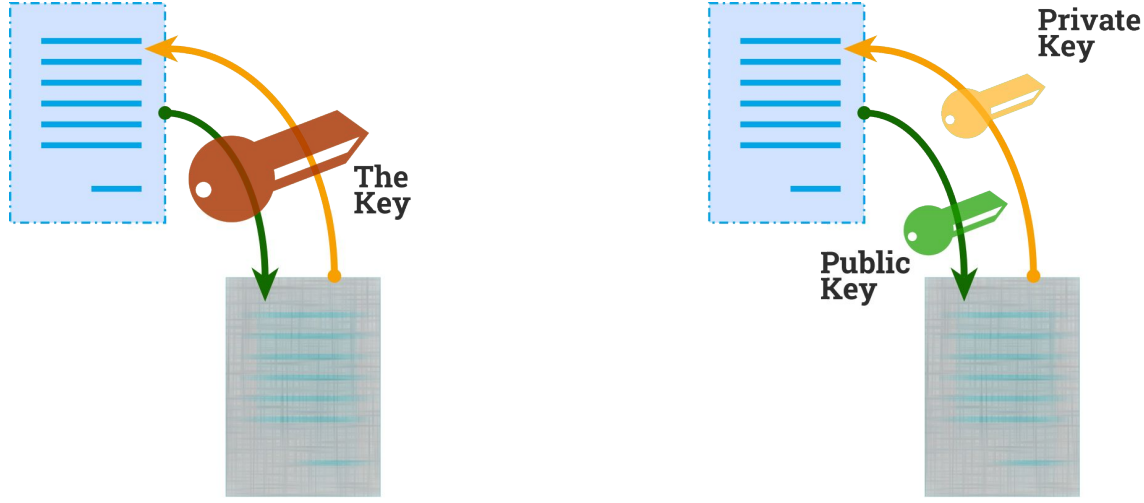


Symmetric encryption



Asymmetric encryption

SSH keys - asymmetric encryption



Symmetric encryption **Asymmetric encryption**

ecdsa-sha2-nistp256
AAAAE2VjZHNhLXNoYTItbmlzdHAyNTYAAAAIbmlzdHAyNTYAAABBBJHT0aW+Q2bjkfvVM7kN2ucaf8eltoaIV6V0CDBgpWF+dKV1HKkvGWWfjtYMOdLDUKlAzJc5Q6bKs6bkYiqqtUQ=

UNIVERSITÄT BONN

Introduction to Virtual Machines (VM)

- high performance computing => server farm
- de.NBI offers VMs (VM = virtual machine)
  - running Linux
  - access through SSH

- de.NBI connected LifeScience AAI Account needed

- Tasks:
  1. Register: https://cloud.denbi.de/wiki/registration/
  2. Apply for the course de.NBI project: https://signup.aai.lifescience-ri.eu/fed/registrar/?vo=denbi&group=pbbcourse20251
  3. create SSH key pair in de.NBI cloud portal and confirm unix user name

UNIVERSITÄT BONN

SSH command and CLI

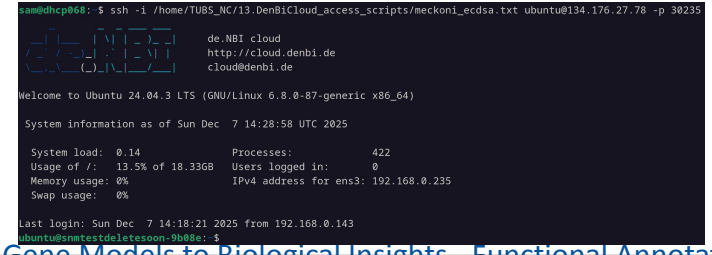
https://github.com/bpucker/teaching/tree/master/deNBI2025

@PuckerLab

- SSH key might need the correct permissions
- needed information:
  - username
  - ip address/domain
  - port number (usually optional but mandatory for de.NBI VMs)
  - ssh private key
  - exemplary commands:

ssh -i /path/to/key username@ip-address -p port-number

ssh -i C:\Users\User\de.NBI_course\key.txt ubuntu@134.176.27.78 -p 23457



UNIVERSITÄT BONN

BUSCO - evaluating structural annotation

- Good structural annotation needed before functional annotation

UNIVERSITÄT BONN

BUSCO - evaluating structural annotation

- Good structural annotation needed before functional annotation

- BUSCO = Benchmarking Universal Single-Copy Orthologs

UNIVERSITÄT BONN

BUSCO - evaluating structural annotation

- Good structural annotation needed before functional annotation
- BUSCO = Benchmarking Universal Single-Copy Orthologs
- Universal genes -> expected to be in the genome
- Singe-Copy -> expected to be present only once per haplophase

UNIVERSITÄT BONN

BUSCO - evaluating structural annotation

- Good structural annotation needed before functional annotation
- BUSCO = Benchmarking Universal Single-Copy Orthologs
- Universal genes -> expected to be in the genome
- Single-Copy -> expected to be present only once per haplophase
- if all BUSCO genes are present in the assembly => good assembly
- some might be natively missing due to gene loss in evolution

UNIVERSITÄT BONN BUSCO

- exemplary commands:

docker container name (will be automatically downloaded)

list datasets:

```
sudo docker run --rm -v /vol/data:/vol/data ezlabgva/busco:v6.0.0_cv1 busco --list-datasets
```

command that will be run in the docker container starts here

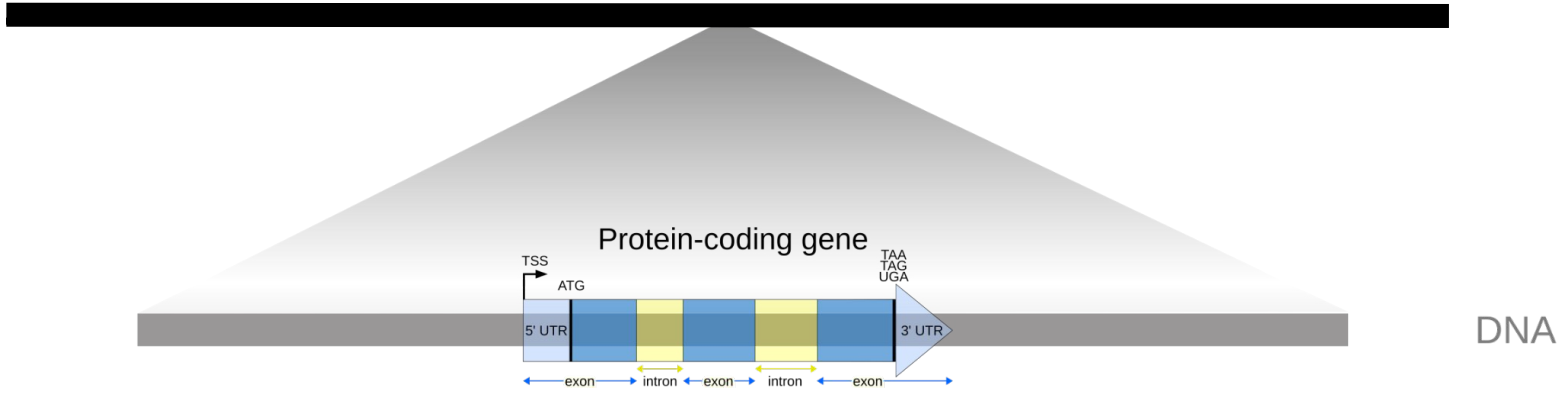
important for the docker container to access files

start the docker container
input file
mode and lineage-dataset
output file

```
sudo docker run --rm -v /vol/data:/vol/data ezlabgva/busco:v6.0.0_cv1 \
busco -i /vol/data/user/protein.fasta \
-m proteins --cpu 13 -l brassicales_odb12 \
--out_path /vol/data/user/BUSCO_pep/ -o busco_run_id --opt-out-run-stats
```

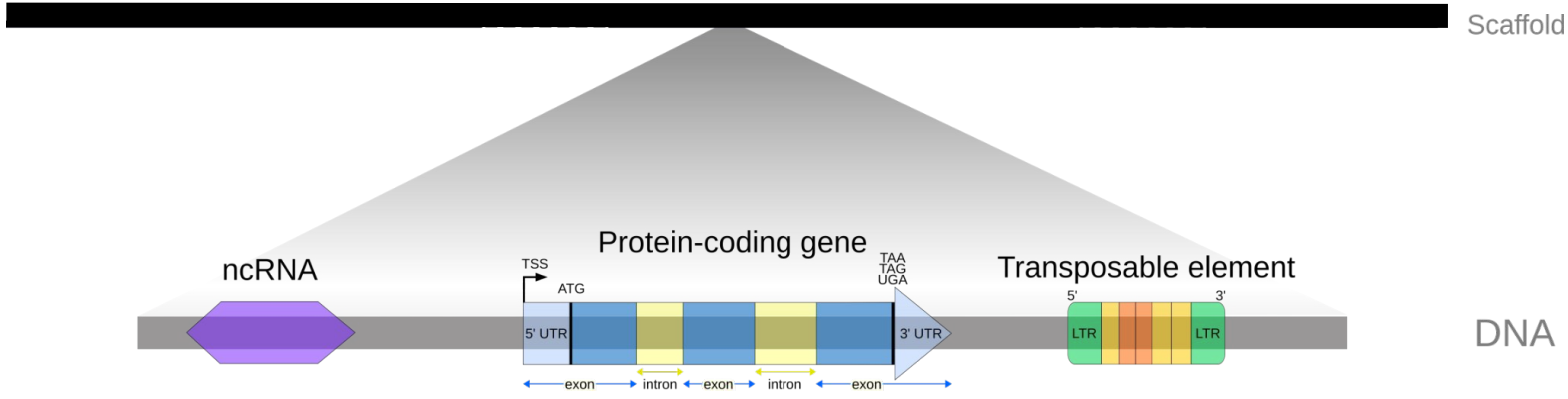
Structural annotation:

finding genes in a genome



Structural annotation:

finding genes in a genome

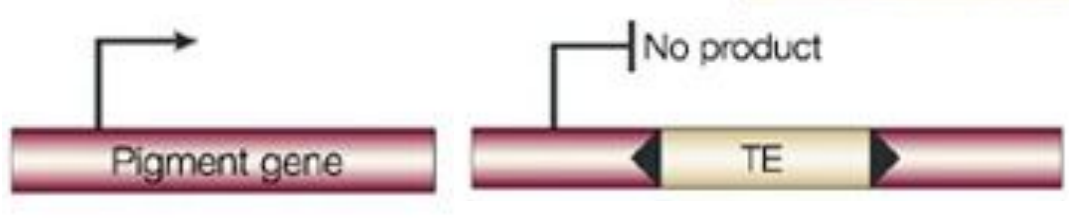


Not just protein-coding genes, but also non-coding DNA like Transposable element (TEs) and non-coding RNAs (ncRNAs)

In plants, only ~1-10% protein-coding genes

Transposable elements

- Discovered by Barbara McClintock in maize
- TEs responsible for large plant genome sizes
- Based on where they land, they dictate the regulation and impact



Feschotte et al, 2002: 10.1038/nrg793
Jiang et al, 2022: 10.1186/s12864-023-09272-x

Classes of TEs

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F

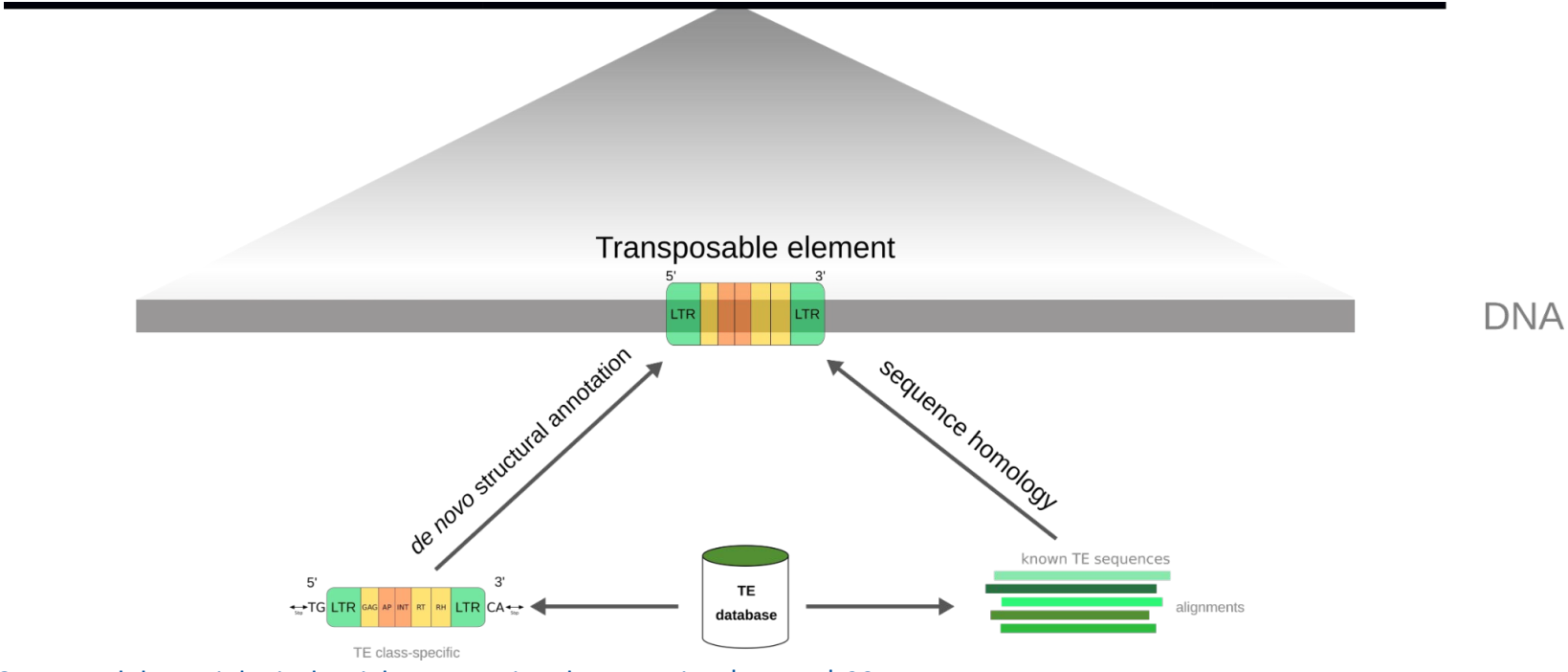
LTR copia

LTR gypsy

LINE L1

Wicker et al, 2007: 10.1038/nrg2165

Annotating TEs



Annotating TEs -Tools

Specific tools

- LTR_FINDER/ LTR_HARVEST ⟶ LTRs
- TIR-LEARNER ⟶ TIRs
- Helitron scanner ⟶ Helitrons

Annotating TEs -Tools

Specific tools

- LTR_FINDER/ LTR_HARVEST ⟶ LTRs
- TIR-LEARNER ⟶ TIRs
- Helitron scanner ⟶ Helitrons

EDTA (Extensive *de novo* TE annotator)

- pipeline combining above tools
- appropriate filtering
- Final TE library

EDTA Installation

Installing Miniconda

cd /vol/data/tools ← #Important not in home

wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh

bash Miniconda3-latest-Linux-x86_64.sh

the above command will interactively ask 4 questions, first press ENTER to continue
Do you accept license terms? [yes|no] →yes
#Miniconda3 will be installed in /home/ubuntu/miniconda3 or specify a different location → /vol/data/tools/Miniconda3
#Proceed with initialization [yes|no] → yes

source ~/.bashrc

conda config --add channels conda-forge

conda update -n base --all

conda install -n base mamba

UNIVERSITÄT BONN

Installing EDTA

Installing EDTA through conda/mamba

mamba create -n EDTA2.2 -c conda-forge -c bioconda -c r annosine2 biopython cd-hit coreutils genericrepeatfinder genometools-genometools glob2 tir-learner ltr_finder_parallel ltr_retriever mdust multiprocess muscle openjdk perl perl-text-soundex r-base r-dplyr regex repeatmodeler r-ggplot2 r-here r-tidyr tesorter samtools bedtools LTR_HARVEST_parallel HelitronScanner

Getting the latest version of EDTA from GitHub

cd /vol/data/tools; git clone https://github.com/oushujun/EDTA.git; cd EDTA; git checkout EDTA2; git branch

Running EDTA

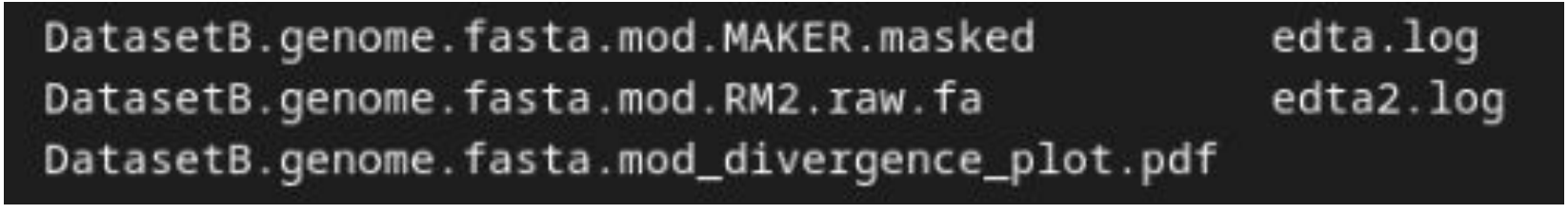
cd /vol/data/user

mkdir EDTA_results; cd EDTA_results

conda activate EDTA2.2

perl /vol/data/tools/EDTA/EDTA.pl --genome /vol/data/DatasetB.genome.fasta --overwrite 1 --anno 1 --sensitive 1 --evaluate 1 --threads 13 >>edta.log 2>&1 &

EDTA- output and results

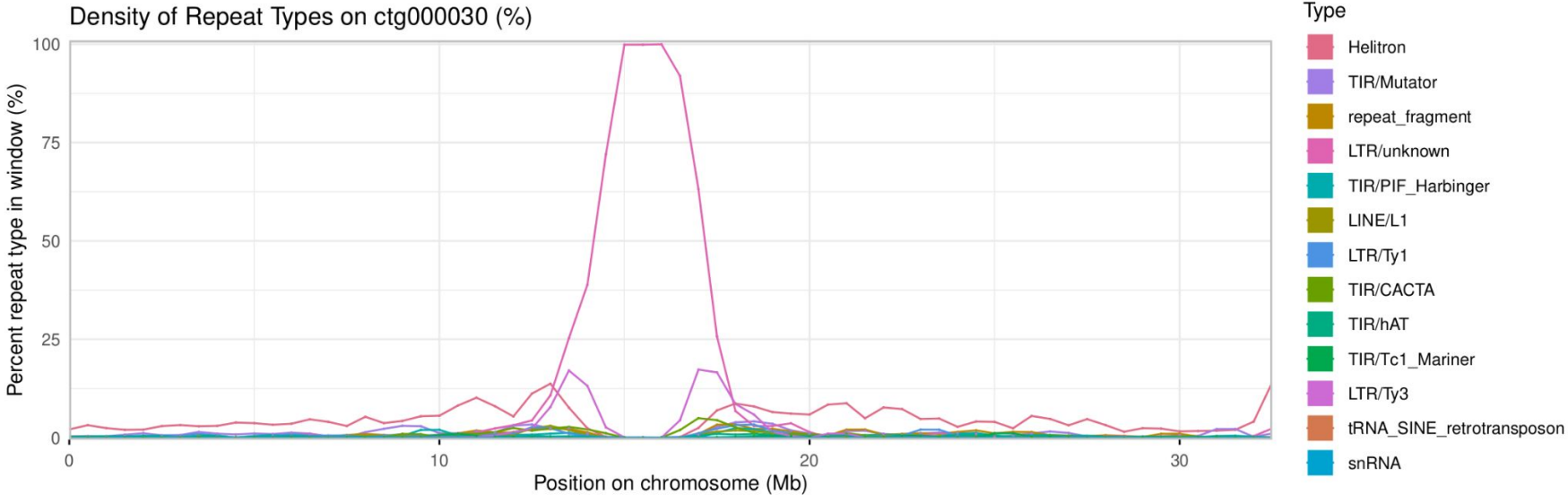


EDTA- output and results

1. *.EDTA.TEanno.sum
2. *.EDTA.TEanno.gff3
3. Density plots for top contigs

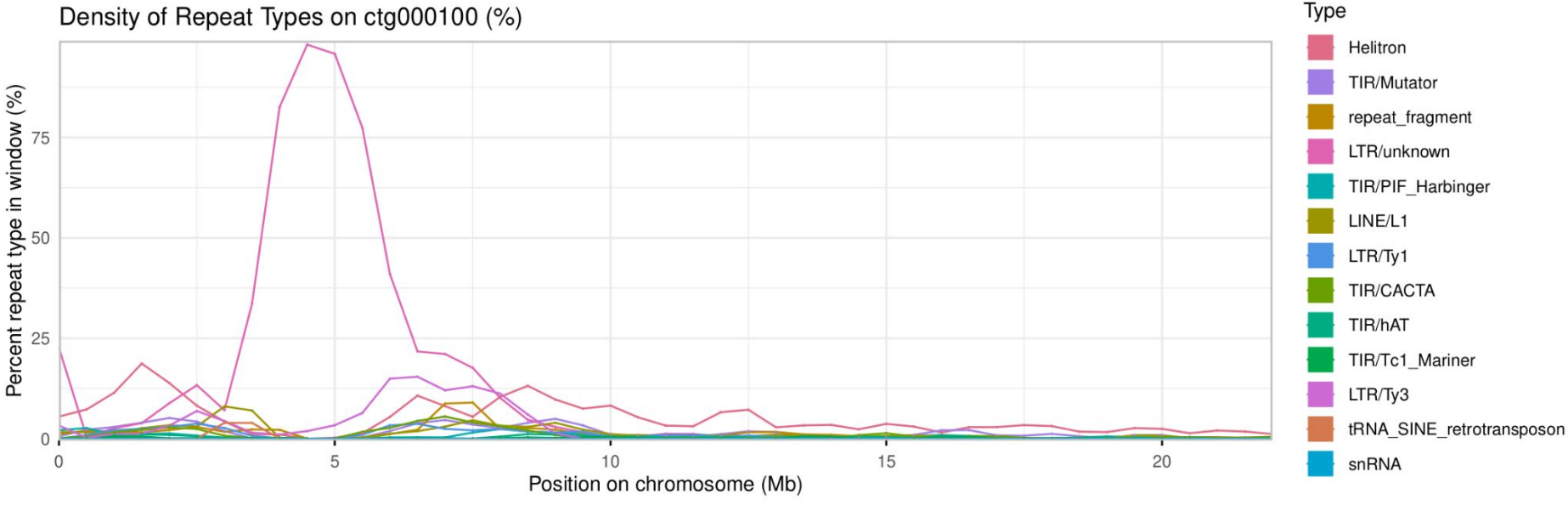
```
Repeat Classes
==============
Total Sequences: 13
Total Length: 134799993 bp
Class                  Count        bpMasked    %masked
=====                  =====        ========     =======
LINE                   --           --           --
    L1                 1239         989157       0.73%
LTR                    --           --           --
    Copia              738          1002959      0.74%
    Gypsy              2081         3195026      2.37%
    unknown            5660         16713457     12.40%
SINE                   --           --           --
    tRNA               143          179011       0.13%
TIR                    --           --           --
    CACTA              1668         918659       0.68%
    Mutator            3170         1767375      1.31%
    PIF_Harbinger      1751         728452       0.54%
    Tc1_Mariner        539          259836       0.19%
    hAT                807          335497       0.25%
nonTIR                 --           --           --
    helitron           11058        5960229      4.42%
repeat_fragment        3879         1580072      1.17%
                       ---------------------------------
    total interspersed 32733        33629730     24.95%

snRNA                  8            859          0.00%
-------------------------------------------------------------
Total                  32741        33630589     24.95%
```

Density plot



UNIVERSITÄT BONN

Density plot



UNIVERSITÄT BONN

End of day 1

Today you learned:

- Accessing de.NBI VMs via SSH
- Working in a CLI environment
- Running BUSCO
- Identifying transposon elements

Tomorrow you will learn:

- tRNA annotation
- Running InterProScan
- KIPEs3, MYB_annotator
- Tree building

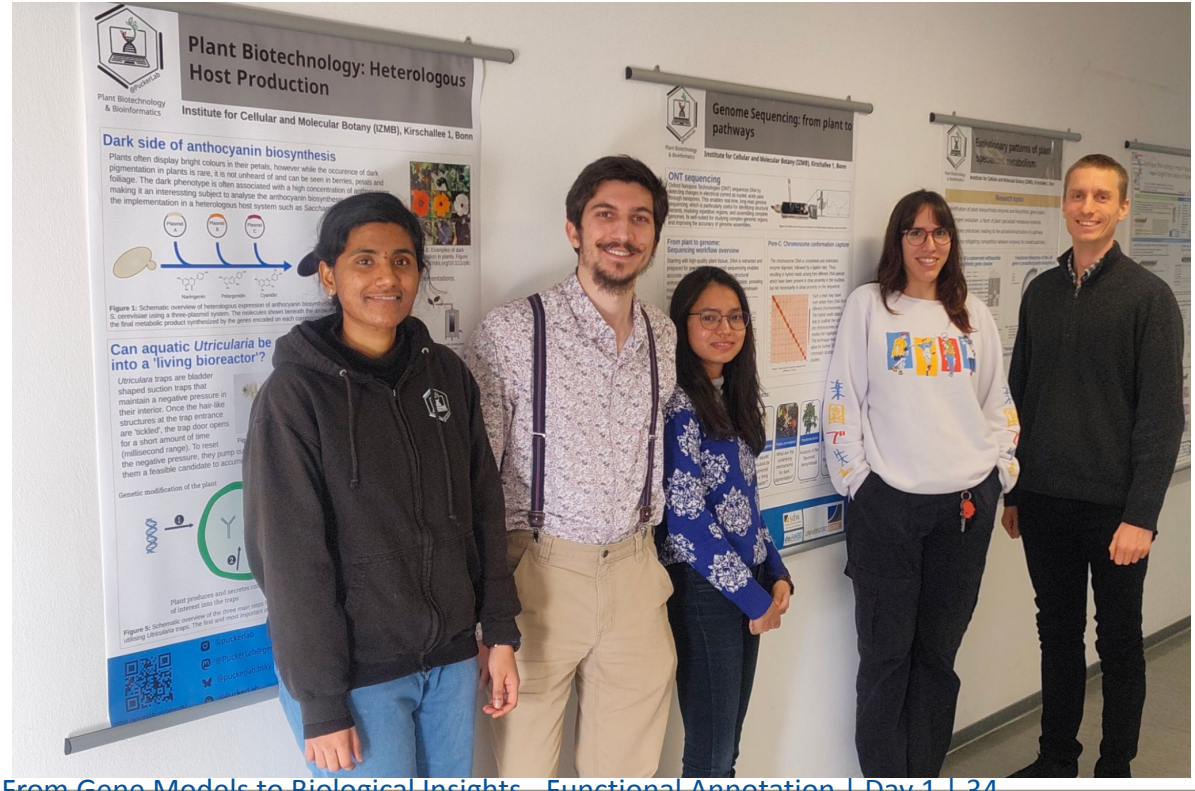
UNIVERSITÄT BONN

Availability of materials

Please find all course materials here:
https://github.com/bpucker/teaching/tree/master/deNBI2025

UNIVERSITÄT BONN

Acknowledgements



@PuckerLab

@PuckerLab

Email: pucker@uni-bonn.de
Web: pbb.uni-bonn.de

Supported by:

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE