1. Sequencing data

DrosEU & Dros-RTEC data:

Automated downloading and processing of 246 paired-end Pool-Seq libraries

DGN data:

Automated downloading and processing of single-individual libraries from 25 populations with ≥ 5 collected individuals and one *D. simulans* sample

2. Metadata

Sample metadata

Collection date
Collector
Catching method
Sample size
Inversion status
SRA number
Sampling coordinates

Environmental metadata

WorldClim
Interpolated bioclimatic trends

Weather-data
Current climatic conditions from weather stations

3. Mapping pipeline

Operation	Program/Script	Output file(s)
Sequence quality check	*FASTQC*	FASTQC output files
Adapter trimming & PE-read merging	*cutdadapt* *bbmerge*	
Mapping against hologenome reference	*bwa mem* *GATK*	Hologenome BAM
D. simulans decontamination	*fix_bam.py*	Decontaminated BAM
Conversion to pileup & gSYNC format	*samtools mpileup* *mpileup2sync.py*	Genome-wide pileup file Genome-wide gSYNC Indel position file Read-depth info
Masking gSYNC file	*MaskSYNC.py*	Masked gSYNC file BED file of masked pos.

4. SNP calling

PoolSNP

SNP calling based on heuristic parameters (common SNPs)

Merging of individual gSYNC files to joint SYNC files

Identifying polymorphic positions based on allele frequency (MAF) and allele counts (MAC)

Creating joint tbi-indexed SYNC/VCF file

Snape

probabilistic SNP calling (population-specific SNPs)

Probabilistic SNP calling based from individual pileup files

Filtering of SNAPE output and conversion to gSYNC file format

Merging individual gSYNC files to joint tbi-indexed SYNC/VCF file

5. Quality control

Basic summary statistics:
Read depth; # of (private) SNPs;
D. simulans contamination

SNP calling parameters:
p_N/p_S ratio and SFS
at different MAF/MAC cutoffs

Comparison of AFs
to Kapun *et al.* 2020 and
Machado *et al.* 2020

Basic biological analyses:
Principal component analyses

6. Resources generated

Extensible pipeline

We provide an **highly automated** and **platform-independent** mapping and SNP calling pipeline which is **highly flexible** and which allows to **easily extend existing datasets**.

Data availability

Using our novel pipeline we combined whole-genome **sequencing data** from **272 world-wide populations** and provide **high-confidence allele frequency data** and corresponding **meta information** in **multiple file formats**

Genome browser

Population genetic estimates and **allele frequency data** can be **visualized** in and **downloaded** from an easy-to-use **genome browser** available at our website: **http://dest.bio**

Demography informative SNP markers

We generated a set of **informative SNP markers**, which allow to assign population samples to predefined **demographic clusters**

Part of Docker pipeline image