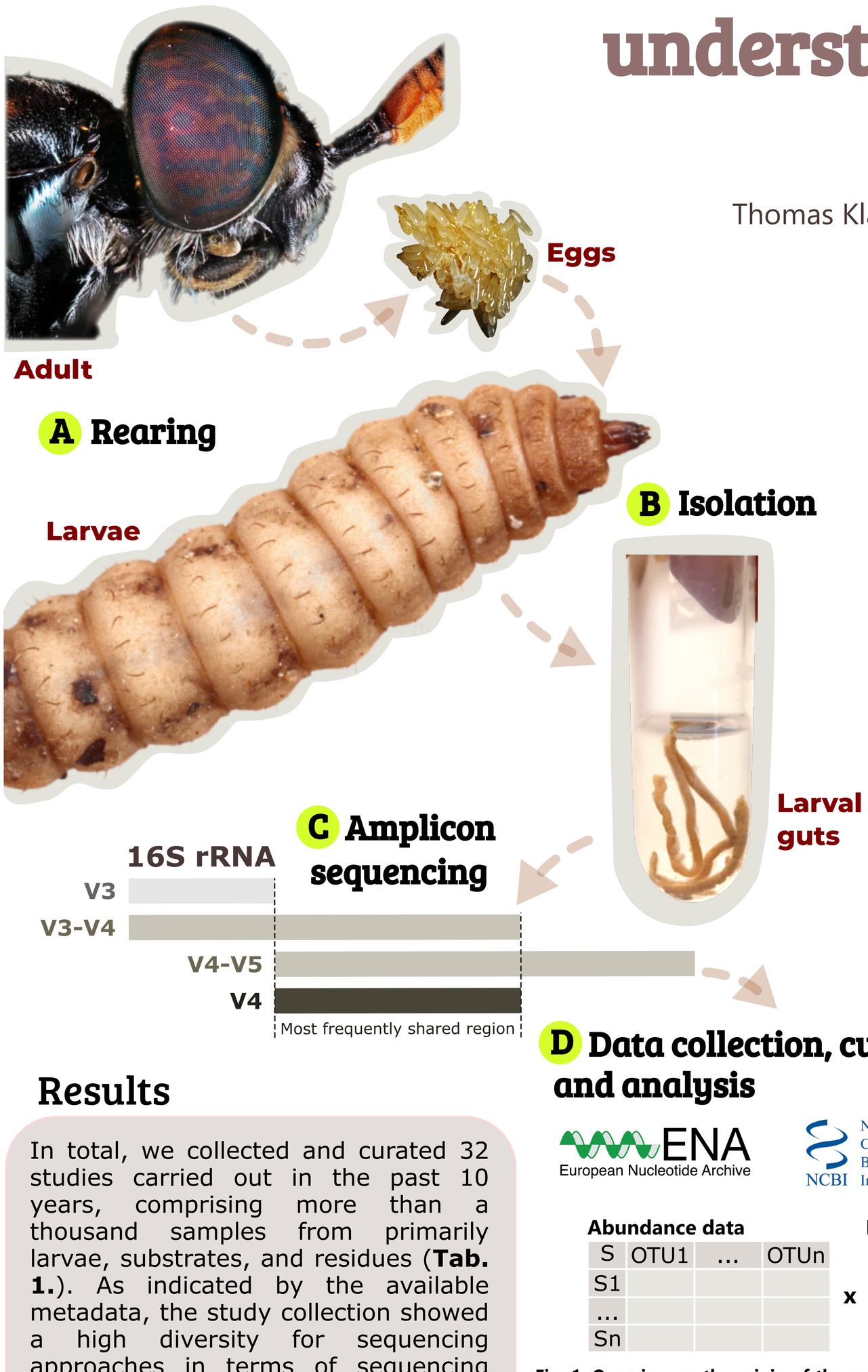
High-performance computing lifts the understanding of insect-based gut microbiomes

Thomas Klammsteiner[a], Heribert Insam[a], Bostjan Murovec[b], Blaz Stres[c,d,e]

[a]Department of Microbiology, University of Innsbruck, Innsbruck, Austria
[b]Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
[c]Department of Animal Science, University of Ljubljana, Ljubljana, Slovenia
[d]Institute of Sanitary Engineering, University of Ljubljana, Ljubljana, Slovenia
[e]Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia



Background

As the flagship species of an emerging insect biotechnology with the goal to convert organic wastes into a source of insect-based protein and fat, the black soldier fly (BSF; *Hermetia illucens*) has awaken popular interest in science and industry. Research showed, a complex network of microbial gut communities strongly contributes to waste degradation processes and shapes the larval environment [1, 2]. However, the lack of a best practice approach in preparing and analyzing raw sequence data affects the informative value and reproducibility between studies. Thus, a comprehensive re-analysis of published data following a standardized protocol in the context of a meta-analysis could drastically improve comparability and enable robust conclusions across studies (**Fig. 1.**). The volume of sequencing data and the iterative preprocessing optimization require for high computational effort, creating the need for parallelization of analysis tools on high-performance computing (HPC) platforms.

Results

In total, we collected and curated 32 studies carried out in the past 10 years, comprising more than a thousand samples from primarily larvae, substrates, and residues (**Tab. 1.**). As indicated by the available metadata, the study collection showed a high diversity for sequencing approaches in terms of sequencing platforms, protocols, depth, and targeted variable regions on the 16S rRNA gene. Moreover, a high variability in experimental designs including sample size, rearing conditions, and composition of fed diets was observed. Consequently, binning and processing datasets with similar sequence features was needed. The V4 region was then isolated for downstream analyses, as it had been covered by 96% of all suitable studies. After isolating the most frequently targeted gene segment, the preprocessed datasets were merged and prepared for taxonomic and functional analysis using GUMPP [3].

Fig. 1. Overview on the origin of the gathered data. A. During their short lifespan, adults lay clutches of up to 600 eggs from which neonate larvae hatch after approx. 3 days. **B.** To study the gut microbiome, the larval gut is in generally dissected before DNA extraction. **C.** The V4 region on the 16S rRNA gene is the most frequently covered region in the collected studies. **D.** (Meta)Data collected from public archives and publications is categorized to construct tables containing environmental and spatial information, which later serve to interpret functional and taxonomic data.

Tab. 1. Unfiltered summary of the collected studies. The values of the listed parameters highlight the high variation between the published studies.

	Parameter	Value
Studies	Number of studies	32
	Number of samples	1008
	Studies with unavailable data	6
	Samples/study (mean ± sd)	28 ± 27
	Countries of origin	20
	Types of feed	22
	Rearing temperature range	20-33 °C
Sequencing	Types of DNA extraction kits	15
	Sequencing platforms	4
	Targeted genetic regions	6
	Reads/sample (mean ± sd)	$9\times10^4 \pm 1\times10^5$

Discussion

Sinking costs with simultaneously increasing sequencing depths are consequences of continuously improving sequencing technologies, condensing in the data presented by microbiome studies. In cross-study meta analyses, strongly varying read number and read quality pose challenges for merging datasets. Especially the drastic increase in data volume while exploring the nature of the data constitute costly and time-consuming computational obstacles. As many sequence analysis tools lack appropriate HPC compatibility and struggle with parallelization, further efforts are needed to increase efficiency and compensate for study-related deficiencies.

Conclusion

Microbiome meta-studies as means to summarize past studies by applying reproducible and well-established protocols have the potential to point out informative cross-study patterns. However, inconsistent or missing metadata and sequence data unavailability hinder such analyses. In addition, to allow for more efficient processing of sequence data on HPC platforms, options for parallelization of analysis tools need improvement.

Funding

This research was funded by a short-term scientific mission grant from EU COST action "Statistical and machine learning techniques in human microbiome studies" (CA18131).

References

[1] Klammsteiner, T., Walter, A., Bogataj, T., Heussler, C.D., Stres, B., Steiner, F.M., Schlick-Steiner, B.C., Insam, H., 2021. **Impact of Processed Food (Canteen and Oil Wastes) on the Development of Black Soldier Fly (*Hermetia illucens*) Larvae and Their Gut Microbiome Functions.** Front. Microbiol. 12. https://doi.org/10.3389/fmicb.2021.619112

[2] Klammsteiner, T., Walter, A., Bogataj, T., Heussler, C.D., Stres, B., Steiner, F.M., Schlick-Steiner, B.C., Arthofer, W., Insam, H., 2020. **The core gut microbiome of black soldier fly (*Hermetia illucens*) larvae raised on low-bioburden diets**. Front Microbiol 11, 993. https://doi.org/10.3389/fmicb.2020.00993

[3] Muorvec, B., Deutsch, L., Stres, B., 2021. **General unified microbiome profiling pipeline (GUMPP) for large scale, streamlined and reproducible analysis of bacterial 16S rRNA data to predicted microbial metagenomes, enzymatic reactions and metabolic pathways**. Metabolites 11, 336. https://doi.org/10.3390/metabo11060336