# Sunbeam: An Extensible Pipeline for Analyzing Metagenomic Sequencing Experiments

Erik Clarke, Louis Taylor, Chunyu Zhao, Jesse Connell, Frederic Bushman, Kyle Bittinger
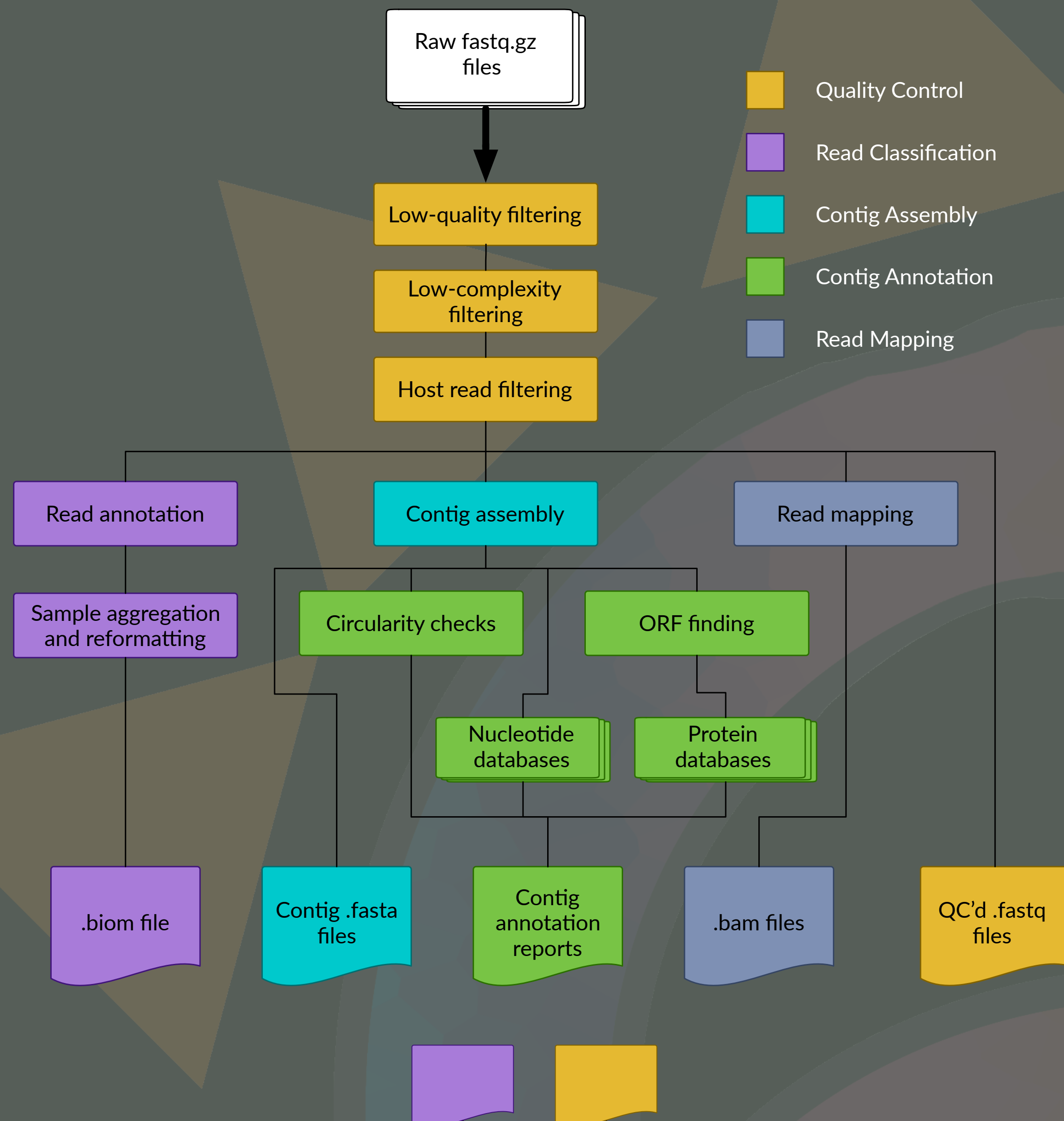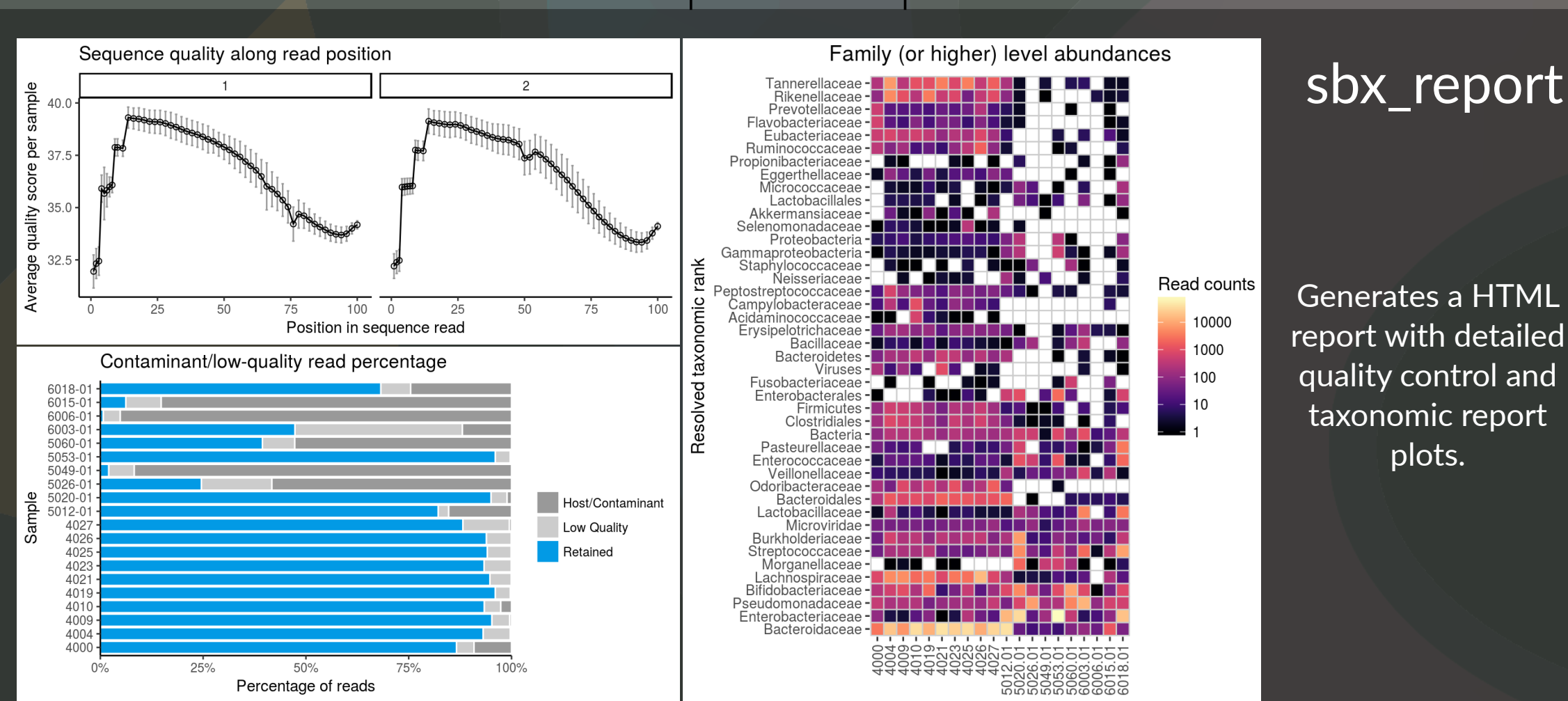
PennCHOP MICROBIOME PROGRAM

Perelman School of Medicine UNIVERSITY OF PENNSYLVANIA

## Workflow



Raw fastq.gz files

- Quality Control
- Read Classification
- Contig Assembly
- Contig Annotation
- Read Mapping

Core pipeline

Low-quality filtering → Low-complexity filtering → Host read filtering

Read annotation — Contig assembly — Read mapping

Sample aggregation and reformatting — Circularity checks — ORF finding

Nucleotide databases — Protein databases

.biom file — Contig .fasta files — Contig annotation reports — .bam files — QC'd .fastq files

Sunbeam extensions



**sbx_report**

Generates a HTML report with detailed quality control and taxonomic report plots.
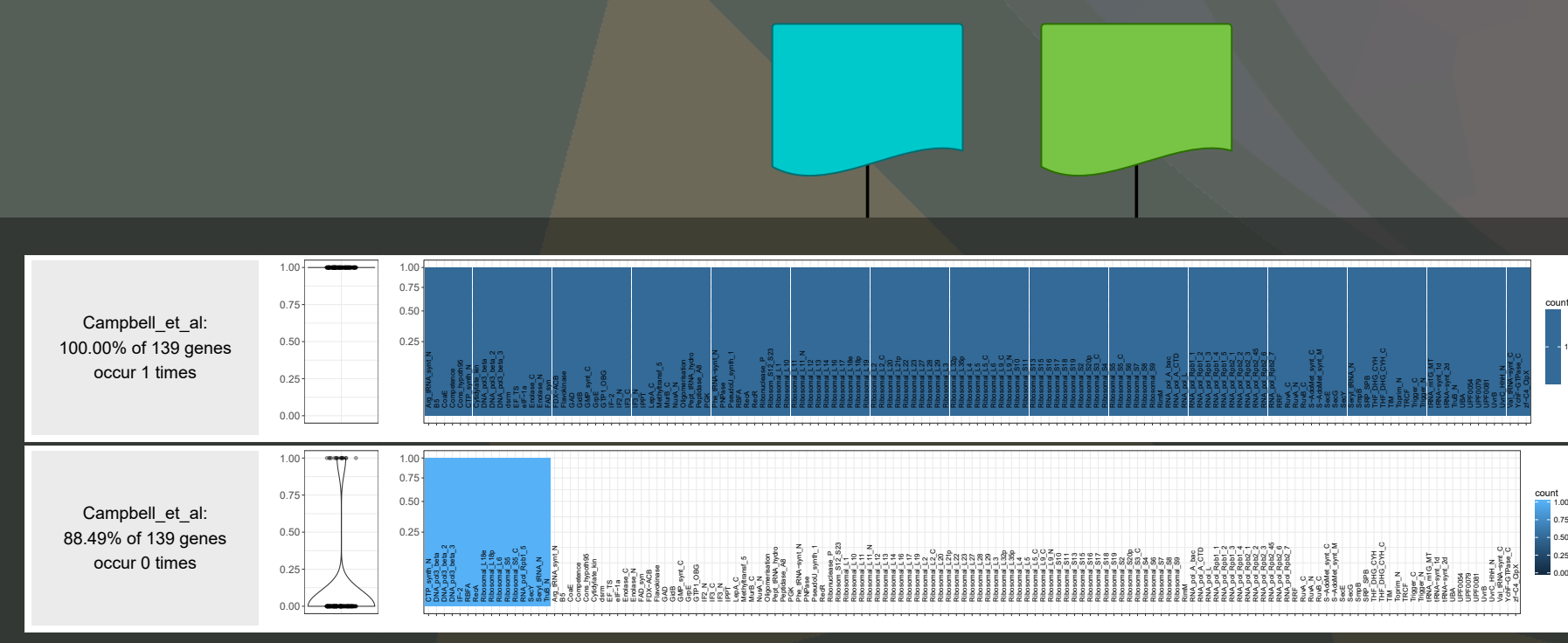


**sbx_anvio**

Interfaces with the Anvi'o visualization platform—here used to visualize single-copy bacterial genes.

**Other prebuilt extensions:**
- **sbx_shortbred**: runs ShortBRED to quantify gene families, ex. antibiotic resistance
- **sbx_metaphlan**: extension to run MetaPhlAn
- **sbx_contigs**: generates coverage plots and summary statistics for contigs by taxon
- **sbx_gene_clusters**: alignment to gene clusters of interest, e.g. bai operon
- **sbx_kaiju**: classify reads with the Kaiju classifier
- and many more!

**sbx_template**

Create your own extension in as few as six lines of code. Then every Sunbeam run includes your custom analyses!

## Abstract

**Background**: Shotgun metagenomic sequencing experiments provide functional and compositional insight into complex microbial communities. To analyze such data, a number of preprocessing and analytical steps must be performed. Many of these steps, such as quality control, adapter trimming, and phylogenetic classification, are common to many sequencing experiments. Other analyses are specific to each study.

**Methods**: Here we introduce Sunbeam, a modular and user-extensible pipeline designed to process metagenomic sequencing data in a consistent and reproducible fashion. Sunbeam performs multiple processing steps common to many metagenomic sequencing experiments including quality control, adapter trimming, host read removal, low-complexity filtering, metagenomic classification, read assembly, and reference genome alignments. Sunbeam also includes a powerful extension framework that enables users to incorporate new analysis or processing steps easily.

**Results**: Sunbeam installs in a single step, has no dependencies other than Linux, doesn't require administrative access, and works on most cluster computing frameworks. Sunbeam is inherently modular and will restart where it left off in case of error. To quickly and accurately filter problematic low-complexity reads in metagenomic data, we also introduce Komplexity, a rapid sequence complexity analysis tool, which identifies low complexity sequences to allow removal. The Sunbeam pipeline is well-documented, regularly updated and in routine use. We also provide a number of pre-built extensions (github.com/sunbeam-labs/).

**Conclusions**: Sunbeam provides an easy-to-use, extensible framework for in-depth analysis of metagenomic sequencing experiments. Sunbeam ensures reproducible and consistent analyses by standardizing post-processing, analytical, and custom steps, and robust removal of problematic, low-complexity reads. Sunbeam is written in Python using the Snakemake workflow management software and is freely available at github.com/sunbeam-labs/sunbeam (GPLv3).

## Key Features

**Modular:** Run pipeline steps in isolation or all together, restart where you left off if anything fails (thanks, Snakemake!)

**Reproducible:** Configuration files are specific to Sunbeam major versions; dependency inter-compatibility through Conda

**Customizable:** Extensions framework for incorporating your own, custom analyses reproducibly and seamlessly

**Easy:** Only requires Linux; install commands fit into a tweet

## Want to learn more? Have questions?

**bioRxiv** THE PREPRINT SERVER FOR BIOLOGY
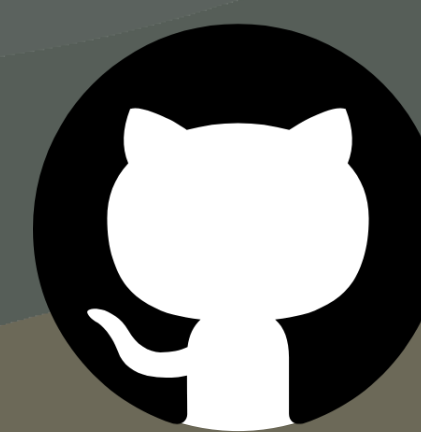
Check out the code for the pipeline and extensions, report issues, get involved in the community:
**github.com/sunbeam-labs**

Read our preprint on bioRxiv:
**doi:10.1101/326363**

Get in touch with us on Twitter:
**@Louviridae (Louis)**
**@pleiotrope (Erik)**
**@zhaocy_1 (Chunyu)**